

## PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-298498

(43)Date of publication of application : 24.10.2000

(51)Int.Cl.

G10L 17/00

G10L 15/04

G10L 15/00

H04N 5/93

(21)Application number : 2000-065101

(71)Applicant : FUJI XEROX CO LTD

(22)Date of filing : 09.03.2000

(72)Inventor : JONATHAN T FOOTE  
WILCOX LYNN D

(30)Priority

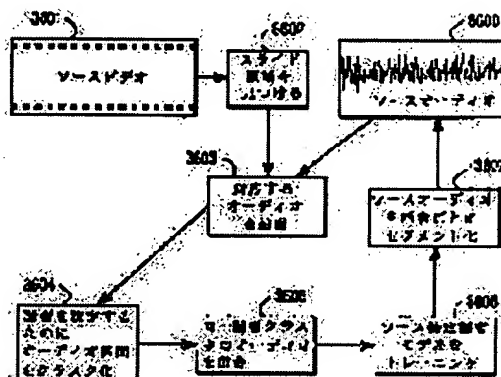
Priority number : 99 266561 Priority date : 11.03.1999 Priority country : US

(54) SEGMENTING METHOD OF AUDIO VISUAL RECORDING SUBSTANCE, COMPUTER STORAGE MEDIUM AND COMPUTER SYSTEM

(57)Abstract:

PROBLEM TO BE SOLVED: To obtain a start of speaker identification of a voice base in order to accurately segmentalize individual presentation through automatic picture recognition by extracting an audio segment corresponding to a video frame segment and applying an acoustic clustering method to the audio segment.

SOLUTION: Source video 3601 is analyzed (3602) to find a slide region. The audio channel of the video 3601 is extracted (3603) for the region of the video 3601 corresponding to the slide segment. The extracted audio segment is clusterized (3604) for every speaker and each of the obtained clusters between audio segments is considered to be based on a single speaker. Audio segments of the same speaker cluster are combined (3605) and a source specific speaker model is trained for each combined audio segment (3606). The audio channel of the source video 3601 is segmentalized for every speaker by speaker recognition (3607).



## LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2000 Japan Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2000-298498

(P2000-298498A)

(43) 公開日 平成12年10月24日 (2000. 10. 24)

(51) Int.Cl. <sup>7</sup>	識別記号	F I	テーマコード* (参考)
G 1 0 L 17/00		G 1 0 L 3/00	5 4 5 F
15/04			5 1 5 C
15/00			5 4 5 C
H 0 4 N 5/93		H 0 4 N 5/93	5 5 1 G
			E
		審査請求 未請求	請求項の数26 O L (全 36 頁)

(21) 出願番号 特願2000-65101 (P2000-65101)

(22) 出願日 平成12年3月9日 (2000. 3. 9)

(31) 優先権主張番号 2 6 6 5 6 1

(32) 優先日 平成11年3月11日 (1999. 3. 11)

(33) 優先権主張国 米国 (US)

(71) 出願人 000005496

富士ゼロックス株式会社

東京都港区赤坂二丁目17番22号

(72) 発明者 ジョナサン ディー フート

アメリカ合衆国 94025 カリフォルニア

州 メンロ パーク ローレル ストリート 450

(72) 発明者 リン ディー ウィルコックス

アメリカ合衆国 94028 カリフォルニア

州 ポートラ ヴァレイ ホワーキン ロード 45

(74) 代理人 100079049

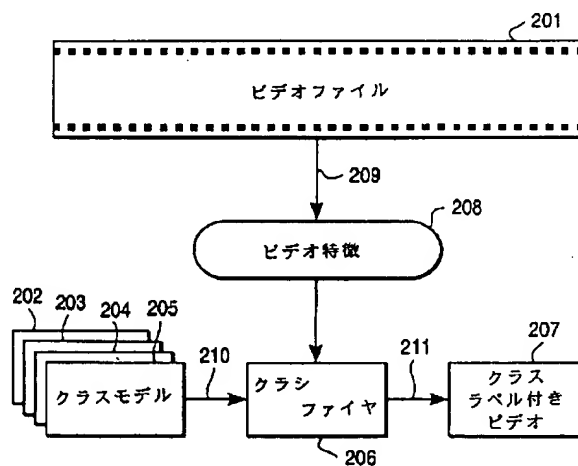
弁理士 中島 淳 (外1名)

(54) 【発明の名称】 オーディオ・ビジュアル記録物をセグメント化する方法およびコンピュータ記憶媒体、並びにコンピュータシステム

(57) 【要約】 (修正有)

【課題】 1人以上の話者によるスライドプレゼンテーションを含む会議のオーディオビデオ記録物のセグメント化方法の提供。

【解決手段】 セグメントは記録された会議の索引として機能する。システムはプレゼンテーションスライドに対応するビデオの区間を自動的に検出し、ビデオでスライドが表示されている時の区間において話者識別技法によって、誰がしゃべっているか推定する。単一話者に対応する併合されクラスタ化されたオーディオ区間はその後、話者セグメント化システムのトレーニングデータとして使用される。話者識別技法によって、ビデオ全体は、各発表者の話の範囲にもとづき個々のプレゼンテーションにセグメント化される。話者識別システムは、各スライド区間からのオーディオデータでトレーニングされた隠れマルコフモデルの構成を選択的に含む。バイタビリティ割り当てがその後、話者に応じてオーディオをセグメント化する。



## 【特許請求の範囲】

【請求項1】 オーディオ・ビデオ記録物をセグメント化する方法であって、  
所定のビデオ画像クラスに対する類似性を有する1個以上のビデオフレーム区間を識別する工程と、  
前記1個以上のビデオフレーム区間に対応する1個以上のオーディオ区間を抽出する工程と、  
1個以上のオーディオクラスタを生成するために前記1個以上のオーディオ区間に音響クラスタ化方法を適用する工程とを含むことを特徴とする方法。

【請求項2】 請求項1記載の方法であって、1個以上のビデオフレーム区間を識別する前記工程が、  
間引かれたフレームを生成するために前記オーディオ・ビデオ記録物のビデオ部分を時間的および空間的に間引く工程と、  
間引かれたフレームの各々について、  
変換マトリックスを生成するために前記間引かれたフレームを変換する工程と、  
前記変換マトリックスから特徴ベクトルを抽出する工程と、  
前記特徴ベクトルおよびビデオ画像クラス統計モデルを用いて前記フレームの類似性を決定する工程とを含むことを特徴とする方法。

【請求項3】 請求項2記載の方法であって、フレームの類似性を測定する前記工程が、  
差分ベクトルを決定するために前記特徴ベクトルからビデオ画像平均ベクトルを減算する工程と、  
前記差分ベクトルの大きさをスレッショルドと比較する工程とを含むことを特徴とする方法。

【請求項4】 請求項3記載の方法であって、前記差分ベクトルの大きさをスレッショルドと比較する工程が、  
前記差分ベクトルの大きさを、前記画像クラス統計モデルに係る標準偏差の所定の倍数と比較する工程を含むことを特徴とする方法。

【請求項5】 請求項1記載の方法であって、所定のビデオ画像クラスに対する類似性を有する1個以上のビデオフレーム区間を識別する前記工程が、  
所定の時間間隔より長いスライド区間に対応するビデオフレーム区間を見つける工程を含むことを特徴とする方法。

【請求項6】 請求項1記載の方法であって、音響クラスタ化方法を適用する前記工程が、  
各オーディオ区間を平均ベクトルによってパラメータ化する工程と、  
各オーディオ区間に対応する平均ベクトルの間のユークリッド距離に集塊クラスタ化法を適用する工程とを含むことを特徴とする方法。

【請求項7】 請求項1記載の方法であって、  
併合オーディオ区間を生成するために同一のオーディオクラスタ内のオーディオ区間を併合する工程と、

前記併合オーディオ区間のソース特定話者モデルをトレーニングする工程とをさらに含むことを特徴とする方法。

【請求項8】 請求項7記載の方法であって、  
各話者を識別するために話者による前記オーディオ・ビデオ記録物を前記ソース特定話者モデルによってセグメント化する工程とをさらに含むことを特徴とする方法。

【請求項9】 請求項7記載の方法であって、  
前記併合オーディオ区間および前記ソース特定話者モデルによって指示される話者シーケンスによって話者遷移モデルを作成する工程と、  
前記話者遷移モデルによって前記オーディオ・ビデオ記録物をセグメント化する工程とをさらに含むことを特徴とする方法。

【請求項10】 オーディオ・ビデオ記録物をセグメント化する方法であって、  
第1の所定のビデオ画像クラスに対する類似性を有する1個以上の第1のビデオフレーム区間を識別する工程と、  
第2の所定のビデオ画像クラスに対する類似性を有する1個以上の第2のビデオフレーム区間を識別する工程と、  
前記1個以上の第1のビデオフレーム区間に対応する1個以上の第1のオーディオ区間を抽出する工程と、  
前記1個以上の第2のビデオフレーム区間に対応する1個以上の第2のオーディオ区間を抽出する工程とを含むことを特徴とする方法。

【請求項11】 請求項10記載の方法であって、  
第1の併合オーディオ区間を生成するために前記1個以上の第1のオーディオ区間を併合する工程と、  
第2の併合オーディオ区間を生成するために前記1個以上の第2のオーディオ区間を併合する工程と、  
前記第1の併合オーディオ区間の第1のソース特定話者モデルをトレーニングする工程と、  
前記第2の併合オーディオ区間の第2のソース特定話者モデルをトレーニングする工程とをさらに含むことを特徴とする方法。

【請求項12】 請求項11記載の方法であって、  
前記第1のソース特定話者モデルによって前記オーディオ・ビデオ記録物の1個以上の第1のセグメントを識別する工程と、  
前記第2のソース特定話者モデルによって前記オーディオ・ビデオ記録物の1個以上の第2のセグメントを識別する工程とをさらに含むことを特徴とする方法。

【請求項13】 コンピュータ可読記憶媒体であって、  
前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、  
所定のビデオ画像クラスに対する類似性を有する1個以上のビデオフレーム区間を識別する工程と、

前記 1 個以上のビデオフレーム区間に対応する 1 個以上のオーディオ区間を抽出する工程と、

1 個以上のオーディオクラスタを生成するために前記 1 個以上のオーディオ区間に音響クラスタ化方法を適用する工程とを含む、オーディオ・ビデオ記録物をセグメント化する方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含むことを特徴とするコンピュータ可読記憶媒体。

【請求項 14】 コンピュータ可読記憶媒体であって、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、1 個以上のビデオフレーム区間を識別する前記工程が、

間引かれたフレームを生成するために前記オーディオ・ビデオ記録物のビデオ部分を時間的および空間的に間引く工程と、

間引かれたフレームの各々について、

変換マトリックスを生成するために前記間引かれたフレームを変換する工程と、

前記変換マトリックスから特徴ベクトルを抽出する工程と、

前記特徴ベクトルおよびビデオ画像クラス統計モデルを用いて前記フレームの類似性を決定する工程とを含む方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含むことを特徴とする請求項 13 記載のコンピュータ可読記憶媒体。

【請求項 15】 コンピュータ可読記憶媒体であって、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、フレームの類似性を測定する前記工程が、

差分ベクトルを決定するために前記特徴ベクトルからビデオ画像平均ベクトルを減算する工程と、

前記差分ベクトルの大きさを前記画像クラス統計モデルに関係する標準偏差の所定の倍数と比較する工程とを含む方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含むことを特徴とする請求項 14 記載のコンピュータ可読記憶媒体。

【請求項 16】 コンピュータ可読記憶媒体であって、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、所定のビデオ画像クラスに対する類似性を有する 1 個以上のビデオフレーム区間を識別する前記工程が、

所定の時間間隔より長いスライド区間に対応するビデオフレーム区間を見つける工程を含む方法をコンピュータが実行するようにプログラムするものである、前記コン

ピュータ可読プログラムコードを含むことを特徴とする請求項 13 記載のコンピュータ可読記憶媒体。

【請求項 17】 コンピュータ可読記憶媒体であって、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、音響クラスタ化方法を適用する前記工程が、

各オーディオ区間をメル周波数ケプストラル係数平均ベクトルによってパラメータ化する工程と、

各オーディオ区間に対応するメル周波数ケプストラル係数平均ベクトルの間のユークリッド距離に集塊クラスタ化法を適用する工程とを含む方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含むことを特徴とする請求項 13 記載のコンピュータ可読記憶媒体。

【請求項 18】 コンピュータ可読記憶媒体であって、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、

併合オーディオ区間を生成するために同一のオーディオクラスタ内のオーディオ区間を併合する工程と、

前記併合オーディオ区間のソース特定話者モデルをトレーニングする工程とをさらに含む方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含むことを特徴とする請求項 13 記載のコンピュータ可読記憶媒体。

【請求項 19】 コンピュータ可読記憶媒体であって、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、

各話者を識別するために話者による前記オーディオ・ビデオ記録物を前記ソース特定話者モデルによってセグメント化する工程とをさらに含む方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含むことを特徴とする請求項 18 記載のコンピュータ可読記憶媒体。

【請求項 20】 コンピュータ可読記憶媒体であって、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、

前記併合オーディオ区間および前記ソース特定話者モデルによって指示される話者シーケンスによって話者遷移モデルを作成する工程と、

前記話者遷移モデルによって前記オーディオ・ビデオ記録物をセグメント化する工程とをさらに含む方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含むことを特徴とする請求項 18 記載のコンピュータ可読記憶媒体。

【請求項 21】 コンピュータ可読記憶媒体であって、前記コンピュータ可読記憶媒体に記憶されたコンピュー

タ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、  
前記話者遷移モデルが、各話者単位がソース特定話者モデルおよびフィルタモデルを含む、一連の話者単位を含むものである方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含むことを特徴とする請求項20記載のコンピュータ可読記憶媒体。

【請求項22】 コンピュータ可読記憶媒体であって、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、オーディオ・ビデオ記録物をセグメント化する方法であり、

第1の所定のビデオ画像クラスに対する類似性を有する1個以上の第1のビデオフレーム区間を識別する工程と、

第2の所定のビデオ画像クラスに対する類似性を有する1個以上の第2のビデオフレーム区間を識別する工程と、

前記1個以上の第1のビデオフレーム区間に対応する1個以上の第1のオーディオ区間を抽出する工程と、  
前記1個以上の第2のビデオフレーム区間に対応する1個以上の第2のオーディオ区間を抽出する工程とを含む方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含むことを特徴とするコンピュータ可読記憶媒体。

【請求項23】 コンピュータ可読記憶媒体であって、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、

第1の併合オーディオ区間を生成するために前記1個以上の第1のオーディオ区間を併合する工程と、

第2の併合オーディオ区間を生成するために前記1個以上の第2のオーディオ区間を併合する工程と、

前記第1の併合オーディオ区間の第1のソース特定話者モデルをトレーニングする工程と、

前記第2の併合オーディオ区間の第2のソース特定話者モデルをトレーニングする工程とをさらに含む方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含むことを特徴とする請求項22記載のコンピュータ可読記憶媒体。

【請求項24】 コンピュータ可読記憶媒体であって、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、

前記第1のソース特定話者モデルによって前記オーディオ・ビデオ記録物の1個以上の第1のセグメントを識別する工程と、

前記第2のソース特定話者モデルによって前記オーディ

オ・ビデオ記録物の1個以上の第2のセグメントを識別する工程とをさらに含む方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含むことを特徴とする請求項23記載のコンピュータ可読記憶媒体。

【請求項25】 コンピュータシステムであって、プロセッサと、

プロセッサ可読記憶媒体に記憶されたプロセッサ可読プログラムコードを有するプロセッサ可読記憶媒体であり、前記プロセッサ可読プログラムコードは、

所定のビデオ画像クラスに対する類似性を有する1個以上のビデオフレーム区間を識別する工程と、

前記1個以上のビデオフレーム区間に対応する1個以上のオーディオ区間を抽出する工程と、

1個以上のオーディオクラスタを生成するために前記1個以上のオーディオ区間に音響クラスタ化方法を適用する工程とを含む、オーディオ・ビデオ記録物をセグメント化する方法を前記コンピュータシステムが実行するようにプログラムするものである、前記プロセッサ可読記憶媒体とを含むことを特徴とするコンピュータシステム。

【請求項26】 コンピュータシステムであって、プロセッサと、

プロセッサ可読記憶媒体に記憶されたプロセッサ可読プログラムコードを有するプロセッサ可読記憶媒体であり、前記プロセッサ可読プログラムコードは、オーディオ・ビデオ記録物をセグメント化する方法であり、

第1の所定のビデオ画像クラスに対する類似性を有する1個以上の第1のビデオフレーム区間を識別する工程と、

第2の所定のビデオ画像クラスに対する類似性を有する1個以上の第2のビデオフレーム区間を識別する工程と、

前記1個以上の第1のビデオフレーム区間に対応する1個以上の第1のオーディオ区間を抽出する工程と、

前記1個以上の第2のビデオフレーム区間に対応する1個以上の第2のオーディオ区間を抽出する工程とを含む方法を前記コンピュータシステムが実行するようにプログラムするものである、前記プロセッサ可読記憶媒体とを含むことを特徴とするコンピュータシステム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 本発明は、内容に従って記録物に自動的に索引づけを行うためのオーディオ・ビジュアル記録物の処理の分野に関する。詳細には、本発明は、記録された会議における個々の口頭でのプレゼンテーションに対応するセグメントを発見する分野に関する。

【0002】

【従来の技術】 従来の方式はオーディオのみのセグメン

ト化に関するものであり、ビデオチャネルはまったく利用されなかった。話者のクラスタ化のための初期データを提供するために均一時間窓(uniform-duration window)を使用することが試みられている。これは、任意の時間の短い窓だけが初期クラスタ化に使用され得るので、初期セグメント化に伴う問題につながる。窓が長すぎると、複数の話者を捕捉する確率が高まり、しかし、窓が短すぎると良好なクラスタ化には不十分なデータしか得られない。補助的な糸口がない場合、窓はしばしば話者の変化を重ね合わせ、それらをクラスタ化にいつそう役立たないものにする。最も従来のセグメント化作業も主としてオーディオにもとづいており、例えば、会議のセグメント化はクローズトリーキングラベルマイクロフォンによる音声認識を使用する。

#### 【0003】

【発明が解決しようとする課題】多くの会議、例えば毎週のスタッフ会議では、1人以上の話者によるスライドプレゼンテーションが含まれる。こうした会議は、以後の再検討および再使用のためにオーディオ・ビジュアル記録媒体に記録されることが多い。そのような会議の内容のブラウジングおよび検索のために、記録された会議の範囲内での各個人の口頭プレゼンテーションの時間の範囲、例えば開始および終了時刻を突き止めることは有益である。

#### 【0004】

【課題を解決するための手段】本発明の第1の態様は、オーディオ・ビデオ記録物をセグメント化する方法であって、該方法は、所定のビデオ画像クラスに対する類似性を有する1個以上のビデオフレーム区間を識別する工程と、前記1個以上のビデオフレーム区間に対応する1個以上のオーディオ区間を抽出する工程と、1個以上のオーディオクラスタを生成するために前記1個以上のオーディオ区間に音響クラスタ化方法を適用する工程とを含む。本発明の第2の態様は、第1の態様において、1個以上のビデオフレーム区間を識別する前記工程が、間引かれたフレームを生成するために前記オーディオ・ビデオ記録物のビデオ部分を時間的および空間的に間引く工程と、間引かれたフレームの各々について、変換マトリックスを生成するために前記間引かれたフレームを変換する工程と、前記変換マトリックスから特徴ベクトルを抽出する工程と、前記特徴ベクトルおよびビデオ画像クラス統計モデルを用いて前記フレームの類似性を決定する工程とを含む。本発明の第3の態様は、第2の態様において、フレームの類似性を測定する前記工程が、差分ベクトルを決定するために前記特徴ベクトルからビデオ画像平均ベクトルを減算する工程と、前記差分ベクトルの大きさをスレッシュホールドと比較する工程とを含む。本発明の第4の態様は、第3の態様において、前記差分ベクトルの大きさをスレッシュホールドと比較する工程が、

前記差分ベクトルの大きさを、前記画像クラス統計モデルに関係する標準偏差の所定の倍数と比較する工程を含む。本発明の第5の態様は、第1の態様において、所定のビデオ画像クラスに対する類似性を有する1個以上のビデオフレーム区間を識別する前記工程が、所定の時間間隔より長いスライド区間に対応するビデオフレーム区間を見つける工程を含む。本発明の第6の態様は、第1の態様において、音響クラスタ化方法を適用する前記工程が、各オーディオ区間を平均ベクトルによってパラメータ化する工程と、各オーディオ区間に対応する平均ベクトルの間のユークリッド距離に集塊クラスタ化法を適用する工程とを含む。本発明の第7の態様は、第6の態様において、前記平均ベクトルがメル周波数ケプストラル係数平均ベクトルである。本発明の第8の態様は、第7の態様において、前記平均ベクトルがフィルタバンクまたはLPC係数平均ベクトルである。本発明の第9の態様は、第1の態様において、併合オーディオ区間を生成するために同一のオーディオクラスタ内のオーディオ区間を併合する工程と、前記併合オーディオ区間のソース特定話者モデルをトレーニングする工程とをさらに含む。本発明の第10の態様は、第9の態様において、各話者を識別するために話者による前記オーディオ・ビデオ記録物を前記ソース特定話者モデルによってセグメント化する工程とをさらに含む。本発明の第11の態様は、第9の態様において、前記併合オーディオ区間および前記ソース特定話者モデルによって指示される話者シーケンスによって話者遷移モデルを作成する工程と、前記話者遷移モデルによって前記オーディオ・ビデオ記録物をセグメント化する工程とをさらに含む。本発明の第12の態様は、第11の態様において、前記話者遷移モデルが、各話者単位がソース特定話者モデルおよびフィルターモデルを含む、一連の話者単位を含む。本発明の第13の態様は、オーディオ・ビデオ記録物をセグメント化する方法であって、第1の所定のビデオ画像クラスに対する類似性を有する1個以上の第1のビデオフレーム区間を識別する工程と、第2の所定のビデオ画像クラスに対する類似性を有する1個以上の第2のビデオフレーム区間を識別する工程と、前記1個以上の第1のビデオフレーム区間に対応する1個以上の第1のオーディオ区間を抽出する工程と、前記1個以上の第2のビデオフレーム区間に対応する1個以上の第2のオーディオ区間を抽出する工程とを含む。本発明の第14の態様は、第13の態様において、第1の併合オーディオ区間を生成するために前記1個以上の第1のオーディオ区間を併合する工程と、第2の併合オーディオ区間を生成するために前記1個以上の第2のオーディオ区間を併合する工程と、前記第1の併合オーディオ区間の第1のソース特定話者モデルをトレーニングする工程と、前記第2の併合オーディオ区間の第2のソース特定話者モデルをトレーニングする工程とをさらに含む。本発明の第15の態様

は、第14の態様において、前記第1のソース特定話者モデルによって前記オーディオ・ビデオ記録物の1個以上の第1のセグメントを識別する工程と、前記第2のソース特定話者モデルによって前記オーディオ・ビデオ記録物の1個以上の第2のセグメントを識別する工程とをさらに含む。本発明の第16の態様は、コンピュータ可読記憶媒体であって、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、所定のビデオ画像クラスに対する類似性を有する1個以上のビデオフレーム区間を識別する工程と、前記1個以上のビデオフレーム区間に対応する1個以上のオーディオ区間を抽出する工程と、1個以上のオーディオクラスタを生成するために前記1個以上のオーディオ区間に音響クラスタ化方法を適用する工程とを含む、オーディオ・ビデオ記録物をセグメント化する方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含む。本発明の第17の態様は、第16の態様において、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、1個以上のビデオフレーム区間を識別する前記工程が、間引かれたフレームを生成するために前記オーディオ・ビデオ記録物のビデオ部分を時間的および空間的に間引く工程と、間引かれたフレームの各々について、変換マトリックスを生成するために前記間引かれたフレームを変換する工程と、前記変換マトリックスから特徴ベクトルを抽出する工程と、前記特徴ベクトルおよびビデオ画像クラス統計モデルを用いて前記フレームの類似性を決定する工程とを含む方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含む。本発明の第18の態様は、第17の態様において、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、フレームの類似性を測定する前記工程が、差分ベクトルを決定するために前記特徴ベクトルからビデオ画像平均ベクトルを減算する工程と、前記差分ベクトルの大きさを前記画像クラス統計モデルに係る標準偏差の所定の倍数と比較する工程とを含む方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含む。本発明の第19の態様は、第16の態様において、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、所定のビデオ画像クラスに対する類似性を有する1個以上のビデオフレーム区間を識別する前記工程が、所定の時間間隔より長いスライド区間に対応するビデオフレーム区間を見つける工程を含む方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含む。

ドを含む。本発明の第20の態様は、第16の態様において、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、音響クラスタ化方法を適用する前記工程が、各オーディオ区間をメル周波数ケプストラル係数平均ベクトルによってパラメータ化する工程と、各オーディオ区間に対応するメル周波数ケプストラル係数平均ベクトルの間のユークリッド距離に集塊クラスタ化方法を適用する工程とを含む方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含む。本発明の第21の態様は、第16の態様において、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、併合オーディオ区間を生成するために同一のオーディオクラスタ内のオーディオ区間を併合する工程と、前記併合オーディオ区間のソース特定話者モデルをトレーニングする工程とをさらに含む方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含む。本発明の第22の態様は、第21の態様において、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、各話者を識別するために話者による前記オーディオ・ビデオ記録物を前記ソース特定話者モデルによってセグメント化する工程とをさらに含む方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含む。本発明の第23の態様は、第21の態様において、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、前記併合オーディオ区間および前記ソース特定話者モデルによって指示される話者シーケンスによって話者遷移モデルを作成する工程と、前記話者遷移モデルによって前記オーディオ・ビデオ記録物をセグメント化する工程とをさらに含む方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含む。本発明の第24の態様は、第23の態様において、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、前記話者遷移モデルが、各話者単位がソース特定話者モデルおよびフィルタモデルを含む、一連の話者単位を含むものである方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含む。本発明の第25の態様は、コンピュータ可読記憶媒体であって、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、オーディオ・ビデオ記録物をセグメント化する方法であり、第1の所定のビデオ画

像クラスに対する類似性を有する1個以上の第1のビデオフレーム区間を識別する工程と、第2の所定のビデオ画像クラスに対する類似性を有する1個以上の第2のビデオフレーム区間を識別する工程と、前記1個以上の第1のビデオフレーム区間に対応する1個以上の第1のオーディオ区間を抽出する工程と、前記1個以上の第2のビデオフレーム区間に対応する1個以上の第2のオーディオ区間を抽出する工程とを含む方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含む。本発明の第26の  
 10 状態は、第25の状態において、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、第1の併合オーディオ区間を生成するために前記1個以上の第1のオーディオ区間を併合する工程と、第2の併合オーディオ区間を生成するために前記1個以上の第2のオーディオ区間を併合する工程と、前記第1の併合オーディオ区間の第1のソース特定話者モデルをトレーニングする工程と、前記第2の併合オーディオ区間の第2のソース特定話者モデルをトレーニングする工程とをさら  
 20 に含む方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含む。本発明の第27の状態は、第26の状態において、前記コンピュータ可読記憶媒体に記憶されたコンピュータ可読プログラムコードであり、前記コンピュータ可読プログラムコードは、前記第1のソース特定話者モデルによって前記オーディオ・ビデオ記録物の1個以上の第1のセグメントを識別する工程と、前記第2のソース特定話者モデルによって前記オーディオ・ビデオ  
 30 記録物の1個以上の第2のセグメントを識別する工程とをさらに含む方法をコンピュータが実行するようにプログラムするものである、前記コンピュータ可読プログラムコードを含む。本発明の第28の状態は、コンピュータシステムであって、プロセッサと、プロセッサ可読記憶媒体に記憶されたプロセッサ可読プログラムコードを有するプロセッサ可読記憶媒体であり、前記プロセッサ可読プログラムコードは、所定のビデオ画像クラスに対する類似性を有する1個以上のビデオフレーム区間を識別する工程と、前記1個以上のビデオフレーム区間に対応する1個以上のオーディオ区間を抽出する工程と、1  
 40 個以上のオーディオクラスを生成するために前記1個以上のオーディオ区間に音響クラス化方法を適用する工程とを含む、オーディオ・ビデオ記録物をセグメント化する方法を前記コンピュータシステムが実行するようにプログラムするものである、前記プロセッサ可読記憶媒体とを含む。本発明の第29の状態は、コンピュータシステムであって、プロセッサと、プロセッサ可読記憶媒体に記憶されたプロセッサ可読プログラムコードを有するプロセッサ可読記憶媒体であり、前記プロセッサ可読プログラムコードは、オーディオ・ビデオ記録物をセ

グメント化する方法であり、第1の所定のビデオ画像クラスに対する類似性を有する1個以上の第1のビデオフレーム区間を識別する工程と、第2の所定のビデオ画像クラスに対する類似性を有する1個以上の第2のビデオフレーム区間を識別する工程と、前記1個以上の第1のビデオフレーム区間に対応する1個以上の第1のオーディオ区間を抽出する工程と、前記1個以上の第2のビデオフレーム区間に対応する1個以上の第2のオーディオ区間を抽出する工程とを含む方法を前記コンピュータシステムが実行するようにプログラムするものである、前記プロセッサ可読記憶媒体とを含む。

【0005】本発明によれば、会議ビデオ記録物における個人のプレゼンテーションを精確にセグメント化するための音声ベースの話者識別の糸口が、自動画像認識により得られる。単一話者のオーディオ区間と関係づけられることが既知のビデオフレーム区間を識別するために、ビデオ変換特徴ベクトルが使用される。オーディオ区間は、オーディオ・ビジュアル記録物のオーディオセグメント化のために話者認識システムをトレーニングするために使用される。

【0006】本発明の好適な実施の形態においては、単一話者の口頭プレゼンテーションはスライドが表示されている区間を含み、特定の話者が各スライドが表示されている時間全体にわたり話すことと仮定される。オーディオ・ビジュアル記録物の単一話者の領域は、スライド画像の拡張された区間を探索することによってビデオ内で識別される。スライド区間は自動的に検出され、それらの領域での音声は、オーディオ話者検出システムをトレーニングするために使用される。単一話者のプレゼンテーションは、話者および聴衆のカメラショットを含むこともある。ある話者によるプレゼンテーションは複数のスライド区間にまたがるのが可能なので、そのスライド区間に対応するオーディオ区間は、ビデオにおいてプレゼンテーションを行う話者の数および順番を見つけるためにオーディオ類似性によってクラスタ化される。クラスタ化の後、単一話者から得られた全部のオーディオデータは、そのオーディオ・ビジュアル記録物のオーディオ部分から特定話者を識別するためのソース特定話者モデルをトレーニングするために使用される。その後オーディオは、話者検出システムによってセグメント化され、そのオーディオ・ビジュアル記録物に索引を付けるための一連の単一話者区間が得られる。

【0007】あるいはまた、スライド以外の単一顔面検出といったビデオ画像クラスのメンバーを探索するか、または、スライドを検出する代わりに演壇の正面に立っている人物を検出するビデオ分析が使用され、単一話者に由来するオーディオに関する区間を検出する。単一話者に関係づけられることが既知のビデオにおけるいずれかの検出可能な特徴が、本発明に従って使用することができる。一般に、単一話者のオーディオ区間と相関して

いることが既知のビデオ画像クラスが、単一話者に由来するオーディオに関する区間を検出するために使用される。

【0008】代替的な実施の形態では、顔面認識が各話者に対応するフレーム区間を検出する。この実施の形態では、特定の話者の顔面認識はその話者のビデオ区間を当該話者によるオーディオ区間と関係づける。従って、顔面認識が、異なる話者によるオーディオ区間を区別する、好ましい実施の形態のオーディオクラス化方法に取って代わる。顔面認識は、認識されたフレームをある話者によるスピーチに関係づける。例えば、第1および第2の話者に対応する第1および第2のビデオ画像クラスは、それぞれ第1および第2の話者に対応するフレーム区間を検出するために使用される。

【0009】本発明によれば、個々のプレゼンテーションに対応する記録された会議の領域が自動的に見つけられる。プレゼンテーションが突き止められると、その領域情報はビデオの索引づけおよびブラウジングのために使用することができる。会議に関係づけられた進行表がある場合、突き止められたプレゼンテーションには、進行表から得られる情報が自動的にラベルづけされる。これにより、プレゼンテーションは発表者および演題によって容易に見つけることが可能になる。

【0010】本発明の方法は、複数の会議ビデオにまたがるように、また、放送ニュースといった他の領域分野に容易に拡張される。本発明の上述および他の特徴および利益は、発明の詳細な説明において図面を参照してより完全に記述される。

#### 【0011】

【発明の実施の形態】ビデオの要約、ブラウジングおよび検索にとって、どのような種類の画像がそのビデオを構成しているかを知るとは、しばしば有益である。例えば、どのショットが人の顔のクローズアップを含んでいるかを知るとは、ビデオの要約にそれらを含めやすくするために有用である。本発明は、ビデオシーケンスを所定のクラスの集合にセグメント化し分類する方法を含む。ビデオクラスの例には、人々のクローズアップ、群衆シーンおよび「パワーポイント（登録商標）」スライドといったプレゼンテーション資料のショットを含む。分類に使用される特徴は一般的であり、従って、ユーザは任意のクラスタイプを指定できる。

【0012】図1は、本発明に従った方法の実施に適する汎用コンピュータシステム100を例示している。汎用コンピュータシステム100は少なくとも1個のマイクロプロセッサ102を備える。カーソル制御装置105は、マウス、ジョイスティック、一連のボタンまたは、ユーザがディスプレイモニタ104上でのカーソルまたはポインタの位置を制御できるようにする他のいずれかの入力装置によって実現される。汎用コンピュータはまた、ランダムアクセスメモリ107、外部記憶装置

103、ROMメモリ108、キーボード106、モデム110およびグラフィックコプロセッサ109を備えることもある。カーソル制御装置105および/またはキーボード106は、本発明に従ってユーザ入力を受け取るための例示的なユーザインタフェースである。汎用コンピュータ100のこれらの要素の全部は、1つの選択肢においては、各種要素間でデータを転送するための共通バス101によって互いに結合されている。バス101は一般に、データ、アドレスおよび制御の各信号を含む。図1に示す汎用コンピュータ100は、汎用コンピュータ100の要素の全部を一体に結合する単一のデータバス101を備えるが、汎用コンピュータ100の各種要素を接続する単一の通信バス101が存在しなければならぬ必要はまったくない。例えば、マイクロプロセッサ102、RAM 107、ROMメモリ108およびグラフィックコプロセッサ109はデータバスによって結合され、ハードディスク103、モデム110、キーボード106、ディスプレイモニタ104およびカーソル制御装置105は第2のデータバス（図示せず）によって接続される。この場合、第1のデータバス101および第2のデータバス（図示せず）は、双方向バスインタフェース（図示せず）によってリンクされる。あるいはまた、マイクロプロセッサ102およびグラフィックコプロセッサ109といった一部の要素は第1のデータバス101および第2のデータバス（図示せず）の両方と接続され、第1のデータバスと第2のデータバスとの間の通信はマイクロプロセッサ102およびグラフィックコプロセッサ109によって行われる。このように、本発明の方法は、図1に100で示したようなあらゆる汎用コンピュータシステム上で実行可能であり、このコンピュータシステムが本発明の方法を実行し得る唯一のものであるといった制限はまったく存在しないことは明白である。

【0013】図2は、本発明によるビデオの分類を実行する方法におけるデータの流れを示す。ビデオファイル201はビデオ記録物のデジタル表現である。ビデオファイル201は一般にMPEGといった標準デジタルフォーマットで符号化されている。画像クラス統計モデル202~205は、4つの個別の画像クラスに対応する所定のガウス分布を表現している。矢印209は、特徴ベクトル208を抽出するためのビデオファイル201の処理を示す。矢印209において行われる処理は以下の通りである。ビデオファイル201は、MPEGといった標準デジタルフォーマットで符号化されている場合、復号化され、画素の矩形マトリックスに変換される。画素の矩形マトリックスは、下位画像のより小形の矩形マトリックスに簡約化され、この場合、各下位画像はその下位画像に対応する画素から導かれるグレースケール符号を表現する。下位画像の矩形マトリックスに変換が施され、変換係数のマトリックスを生じる。変換

係数のマトリックスから、ビデオ特徴208が、ビデオ分類のためのビデオ集合として示される変換マトリックス内の係数位置にある変換係数として選択される。クラシファイヤ(分類ユニット)206は各ビデオ特徴208を受け取り、それらのビデオ特徴208を画像クラス統計モデル202~205の各々に入力する。この結果、ビデオファイル201の各フレームは、画像クラス統計モデル202~205により表現される画像クラスのいずれかに分類される。ビデオファイル201のフレームに対応するようにクラシファイヤ206によって決定された対応する画像クラスは、クラスのラベル付けされたビデオ207に索引づけられる。このようにして、クラスラベル付けされたビデオ207は、そのフレームが属する画像クラスを示す各フレームに関係づけられた情報を含む。

【0014】図2に示す通り、システムは最初に、ビデオシーケンスから分類のための特徴、例えば離散コサイン変換係数を抽出するが、カラーヒストグラムといった他の特徴を選択的に使用することもできる。認識されるビデオの各フレームのモデルを構築するために、トレーニングデータが使用される。このトレーニングデータは、そのクラスからの単数または複数のビデオシーケンスより構成される。クラスモデルは、ガウス分布または隠れマルコフモデルのどちらか一方にもとづけることができる。未知のビデオからクラスモデルおよび特徴が与えられたと、システムは、そのビデオをセグメント化し、それぞれのクラスに属するセグメントに分類する。

【0015】ガウス分布型クラシファイヤは、クラスモデルを用いて各フレームの尤度(likelihood)を計算する。そのフレームのクラスは最大尤度を有するクラスである。同じクラスラベルを有する隣接フレームは併合されてセグメントを形成する。さらに、その尤度は、各クラスにおける帰属関係の信頼の程度を表示するブラウザにおいて選択的に使用される。隠れマルコフモデルの場合、隠れマルコフモデル状態は異なるビデオクラスに対応する。バイタービ(Viterbi)アルゴリズムが使用される。最大尤度状態シーケンス、従って各フレームのクラスラベルを見つけるため、信頼度のスコアは状態シーケンスの確率から得られる。この隠れマルコフモデルクラシファイヤは、上記のフレームごとのクラシファイヤよりも複雑であるが、セグメントの連続性および順序を強制することによってセグメントを平滑化する役割を果たす。これは、単一フレームのクラス決定の変更を効果的に禁止する。

【0016】各画像またはビデオフレームは、離散コサイン変換またはアダマール変換といった変換を用いて変換される。多くの用途の場合、完全なビデオフレームレートは不要であり、フレームは、選択的に数個のフレームのうちの1個だけが変換されるように時間的に間引かれる。この間引きによって、記憶コストおよび計算時間

は劇的に軽減される。画像圧縮においては、一般に小さな下位ブロックに対して変換が行われるが、ここではフレーム画像全体に変換が適用される。変換されたデータはその後、有意性の低い情報を破棄することによって低減される。これは、切り捨て(truncation)、主成分分析または線形識別解析などといった多数の技法のいずれかによって行われる。この用途の場合、また、経験的に示される通り、主成分分析が良好に作用する。それが特徴次元の相関を分離する傾向があり、従って、データが、後述の通りガウスモデルおよび隠れマルコフモデルの対角共分散仮定によく一致するからである。しかし、最大分散を有する係数を単純に選択することが極めて有効であると判明している。これは、各フレームに関してコンパクトな特徴ベクトル(簡約化された係数)をもたらす。この表現は、類似の画像のフレームが類似の特徴を有するので、分類にとって適切である。

【0017】図3は、本発明による、トレーニングフレーム、トレーニングフレームから得られた平均特徴ベクトルの逆離散コサイン変換およびトレーニングフレームから得られた平均特徴ベクトルの逆アダマール変換を例示する。従って、トレーニングフレーム301~308は、ビデオ画像クラスに関係する一連のトレーニング画像を表す。トレーニング画像301~308によって表現された画像クラスは、英語で「演壇の正面に立つ話者」と説明される。フレーム310は、トレーニングフレーム301~308から抽出された8成分特徴ベクトルにもとづいて計算された平均特徴ベクトルに対応する逆離散コサイン変換を図示している。フレーム310では、ビデオ分類のための特徴集合は10成分特徴集合である。従って、各フレームからの10個の変換係数だけが各トレーニングフレームに関係づけられた特徴ベクトルを構成する。フレーム311は、トレーニングフレーム301~308の各々から抽出された100成分特徴ベクトルにもとづいて計算された平均特徴ベクトルの逆離散コサイン変換を表す。フレーム312は1000成分特徴ベクトルの逆離散コサイン変換である。フレーム312は、逆離散コサイン変換において使用される係数の数が増加しているため、それ自体がフレーム310よりも詳細な表示をしているフレーム311よりもさらに詳細に表示している。

【0018】フレーム320は、トレーニング画像から得られた平均特徴ベクトルの逆アダマール変換を表す。フレーム321は100成分特徴ベクトルに対応する逆アダマール変換を表す。フレーム322は1000成分特徴ベクトルに対応する逆アダマール変換を表す。

【0019】1/2秒間隔で取られたMPEGフレームは、復号化され、64×64グレイスケール強度下位画像に簡約化された。得られたフレーム画像は、離散コサイン変換およびアダマール変換により符号化された。最大分散(順位)を有する係数および最も重要な主成分の

両者が特徴として選択された。ガウスモデルは、1～1000の可変数の次元によってトレーニング集合でトレーニングされた。図3は特徴カテゴリの1つ（figure）のサンプルを示す。このカテゴリは、明るい（白い）背景を背にした人々のクローズアップよりなる。このクラスの画像が、カメラアングル、照明および位置の点で、おそらくは典型的なニュースキャスターの画像よりもいかに大きく変化し得るかに留意されたい。平均および共分散は、最大分散の離散コサイン変換およびアダマール変換の係数によってトレーニングされた。各モデルは、ゼロに設定された破棄係数を有する平均を逆変換することによって画像化されている。共分散は示されていないが、平均がトレーニングデータからの主要な特徴（暗色の中央の人影）を捕捉することは明白である。図3は、少ない数の係数によっても、トレーニングデータにおける主要な形状が、逆変換された時に依然認識可能であることを示している。

【0020】図4は、異なる平均および分散を有する2つの次元ガウス分布を示す。確率曲線401によって\*

$$P(x) = ((2\pi)^{-d/2} |\Sigma_c|^{-1/2}) \exp(-1/2(x-\mu_c)' \Sigma_c^{-1}(x-\mu_c)),$$

【0023】ここで、 $\mu_c$ は平均特徴ベクトル、 $\Sigma_c$ はモデルcに関係するd次元特徴の共分散マトリックスである。式 $(x-\mu_c)'$ は差分ベクトルの変換である。実際には、対角共分散マトリックス、すなわち $\Sigma_c$ の非対角線上成分がゼロであると仮定するのが普通である。これにはいくつかの利点がある。最も重要なことは、自由パラメータ（マトリックス成分）の数をd（d-1）/2からdに減らすことであり、これは問題の次元d（dは100のオーダー）が高い時に重要となる。共分散マトリックスは少数のトレーニングサンプルにより計算される際にしばしば不良条件となるので、これはマトリックスの逆の計算が極めて単純になり、より確固としたものになることを意味する。このようにガウスモデルによって画像を分類するために、必要なクラスの各々について1組のサンプルトレーニング画像が集められ、パラメータベクトル $\mu_c$ および $\Sigma_c$ が計算される。未知の画像xが与えられると、各画像クラスの確率が計算され、その画像は最大尤度モデルによって分類される。あるクラス（トレーニング集合）にとっては対数尤度だけが類似性の有用な測度であり、本発明によるビデオブラウザといった用途において直接使用される。より精緻なモデルは、多数のパラメータおよび混合重みを評価するために期待値最大化アルゴリズムが与えられた時に、ガウス混合を使用することができる。さらなる代替として、ニューラルネットワークまたは他の形式のクラシファイヤが使用される。単一のガウスモデルの場合、 $\mu_c$ および $\Sigma_c$ の計算は、計算法としては容易であり、極めて迅速に行える。単一画像からのモデルのトレーニングの場合、平均ベクトルは画像特徴に設定され、分散ベクトル（対角

\*表現された分布Aは平均 $\mu_A$ を有する。確率曲線402によって表現された分布Bは平均 $\mu_B$ を有する。分布Aから生じるある値Xの確率は、横軸に対する点403の垂直位置である。同様に、分布Bより生じる値Xの確率は、横軸に対する点404の垂直高さである。点403における確率が点404における確率よりも高いので、Xは分布Aから最も得られやすい。図4は次元プロットであり、2つの画像クラスAおよびBならびに1成分特徴集合が与えられた時、図4は、本発明に従って行われるビデオフレームの分類の最大尤度の方法を的確に例示する。

【0021】特徴データが与えられると、ビデオセグメントは統計的にモデル化される。単純な統計モデルは多次元ガウス分布である。ベクトルxが1フレームの特徴を表現すると仮定すると、そのフレームがガウスモデルcによって生成される確率は次式の通りである。

【0022】

【数1】

共分散マトリックス）は全部の画像に対する大域変数の何らかの比に設定される。未知のフレームおよび数個のモデルが与えられた場合、その未知のフレームは、どのモデルが最大確率をもってそのフレームを生じるかによって分類される。

【0024】図5は、本発明によるビデオ分類のための特徴集合を選択する例示的方法を示す。すなわち、図5は、統計モデルのトレーニングのため、および、統計モデルがトレーニングされた際のビデオの類似性測定および分類のために、抽出および分析する係数位置を変換する選択過程を表している。図5に記載した方法は、多数のトレーニング画像に見られるの特性を考慮している。以下に述べる分類法において、特徴集合を最適に選択するために使用されるトレーニング画像は、異なるクラス全部の画像を含む。これは、図5に示す方法が異なるクラスの画像を区別するために特徴の最適な集合を選択するのを助ける。図5に示す方法の代替として、特徴集合で使用する係数位置は、観測されるビデオ特性を全く考慮せずに、図6および8に示すように単に最低頻度係数を選択することによる切り捨てによって選択される。

【0025】V×Hの離散コサイン変換係数位置を考え、そこから特徴集合としてより小さな数dを選択する。図6に示した例ではV=H=8である。より典型的で実際のシナリオではV=H=64であり、従って、選択すべき4096（64×64）個の係数位置がある。最大分散係数を選び出すための1つの代替の方法は、4096×4096共分散マトリックスを計算した後、適切に特徴を選び出すが、必ずしも順番に行う必要はない。簡約化ベクトルの実際の順序は重要ではない

が、一致していなければならない。

【0026】工程501で、平均係数マトリックスが計算される。平均係数マトリックスは、変換が適用される下位画像のマトリックスと同じ行数Vおよび同じ列数Hを有しており、また、結果として得られる変換係数マトリックスとも同数の行および列を有する。平均マトリックスの各位置は、トレーニング画像にある対応する係数の算術平均である。1つの実施の形態では、平均係数マトリックスは、分散マトリックスを計算する過程の予備工程として計算される。別の実施の形態では、平均係数マトリックスの値自体を解析して、特徴集合を選択する。例えば、ある実施の形態では、最大平均値を有する係数位置が特徴集合として選択される。工程502では、分散マトリックスが計算される。分散マトリックスは、平均マトリックスおよび変換マトリックスと同じ行数Vおよび同じ列数Hを有する。分散マトリックス502の各値は、トレーニング画像の変換マトリックスにおける対応する位置の統計分散を表現する。あるいはまた、分散マトリックス502の各値は、標準統計分散以外である「分散」測度を表現するが、それでもやはりそれは変動の測度を表現する。例えば、観測された各係数の平均係数との差の算術平均絶対値は、標準統計分散に使用されるような2乗差の和よりも、「分散」測度として使用できる。

【0027】工程503において、特徴集合が選択される。この特徴集合は、本発明による多様な方法のいずれかによって工程503で選択される。例えば、特徴集合は選択的に、最大平均値を有するd個の係数位置として選択される。あるいはまた、特徴集合は分散マトリックスで最大分散値を有するd個の係数位置として選択される。さらに別の代替法として、特徴集合は、主成分分析または線形識別解析によって選択される。

【0028】最も単純な特徴集合選択法では、特徴集合のd個の係数位置が切り捨てによって選択され、それにより、変換マトリックスの最低頻度係数のみが、トレーニングフレームのいずれかにおけるそれらの位置の実際の係数の値にかかわらず、特徴集合を構成するように選択される。実際、切り捨てによると、最低頻度成分が最も重要であると単純に仮定されるので、いずれのトレーニングフレームもまったく分析される必要はない。

【0029】特徴集合の選択はトレーニング画像の各群について行われる必要はないことに留意しなければならない。一般に、特徴集合は、分類方法において使用される全部のクラスモデルから全部のトレーニング画像を使用する上記の方法のいずれかにもとづいて選択される。例えば、図2のクラスモデル202~205の各々を定義するために使用されるトレーニング画像の全部は、それらのトレーニング画像の全部について平均マトリックスおよび分散マトリックスを計算することによって解析されて、それらのクラスモデルの各々の分類のための最

適な特徴集合を決定する。従って、本発明による分類法における各ビデオ画像クラスについて同じ特徴ベクトルが検索されるように、好ましくは同一の特徴集合が全部のクラスモデルに関して使用される。しかし、本発明による画像クラスの各々について同一の特徴集合が使用されなければならないという必要性はまったくない。これに関して、各画像クラスは、その画像クラスの検出に最適に選択された特徴集合を有することができるが、その画像クラスの対応する確率の計算を行うために各ビデオフレームから異なる特徴ベクトルを抽出しなければならないという演算負荷の増加を伴う。

【0030】図6は、ビデオフレームの離散コサイン変換から得られる変換マトリックスを示す。列1は水平周波数0(従って直流)を表現し、列2は水平周波数 $f_h$ を表現し、そして、列8は水平周波数 $13f_h$ の係数を表す。同様に、行1は垂直周波数0(すなわち直流)の係数を表現し、行2は垂直周波数 $f_v$ を表現する。変換マトリックス600の行8は垂直周波数 $13f_v$ の係数を表す。変換マトリックス600の左上隅の9個の係数はその変換マトリックスの最低周波数係数を表す。ブラケット601および602で囲まれたこれらの9個の係数は、本発明による特徴集合を選択する9係数切り捨て法によって選択される9個の係数位置である。より高周波数の係数は画像の細部を表現するので、それらはあるフレームのビデオ画像クラスを決定するうえでそれほど重要でないことがほとんどである。

【0031】図7は、本発明に従って2個以上の変換マトリックスから計算された分散マトリックスを示す。図8は、本発明に従った切り捨てによって決定された特徴集合800を示す。最低頻度成分に対応する変換マトリックスの9個の係数は、図8に示す通り特徴集合800として選択された。例えば、成分801、802および803は図6に示す変換マトリックス600の行1の最初の3個の係数位置を表し、成分804、805および806は変換マトリックス600の第2の行の最低頻度成分を表し、成分807、808および809は変換マトリックス600の第3の行の最低頻度係数位置を表す。変換マトリックス600の最初の3個の行は変換における最低垂直頻度を表し、従って特徴集合800で指定された9個の成分は切り捨て法に関して適切な選択である。

【0032】図9は、本発明に従って図8に示した特徴集合を有するトレーニングフレームの2個の特徴ベクトルから計算された平均特徴ベクトル900を示す。このように、係数801~809に対応する平均マトリックス(図示せず)の値は平均特徴ベクトル900として記憶される。

【0033】図10は、本発明に従って図8に示した特徴集合を有するトレーニングフレームの2個以上の特徴ベクトルから計算された対角共分散マトリックスを示

す。共分散マトリックスは必ず正方かつ対称である。この共分散は次元 $d \times d$ のマトリックスである。共分散は全部の異なる次元に関する相関を表現する。対角共分散を使用することによって、 $d$ 個の非ゼロ値が存在し、数学演算のためには、それはマトリックスとして扱われなければならないものの、 $d$ 成分のベクトルとしてみなすことができる。対角共分散マトリックス1000の全部の非対角線上成分は、特徴集合における全部の特徴がその特徴集合の他の特徴と統計的に非相関関係にあるという仮定にもとづき、ゼロに設定される。実際、特徴が相

【0034】図11は、本発明の方法による図8に示した特徴集合を有するフレームについて検索された特徴ベクトル1100を示す。このように、特徴ベクトル1100の各成分1101～1109は、変換された画像フレームから得られた実際の変換係数を含む。特徴ベクトル1100は、本発明に従った分類方法においてビデオファイル201から抽出される図2に示したビデオ特徴208の実例である。

【0035】図12は、本発明により2個以上のビデオ画像クラスのいずれかにビデオのフレームを分類する方法を例示している。この方法は工程201に始まり、工程202で、ビデオの第1のフレームが離散コサイン変換またはアダマール変換のいずれか一方によって変換される。工程1203で、特徴工程によって示された位置にある係数に対応する特徴ベクトルが抽出される。工程1204では、特徴ベクトルを生成する各画像クラス統計モデルの尤度または確率が計算される。工程1205で、そのフレームに対応する特徴ベクトルを生成する確率が最も高い画像クラス統計モデルを有する画像クラスが選択される。工程1206で、そのフレームが、工程1205で決定されたそのクラス指定によりラベルづけされる。この工程では、フレームは、将来容易にブラウズまたは検索されるように、そのクラスに従って索引づけがなされる。検査1207は、そのビデオにさらにフレームが存在するかどうか、すなわち、それが分類中のビデオの最後のフレームであるかどうかを判定する。さらにフレームがあれば、分岐1208は方法を次のフレームを変換する工程1202へ戻し、それがビデオの最終フレームである場合には、工程1209は図2に示したクラスのラベルづけされたビデオ207が完了したことを指示する。

【0036】図13は、本発明に従い切り捨て以外の方法によって決定された特徴集合を示す。例えば、主成分分析、最大分散を有する係数の選択または最大平均を有

する係数の選択の内の1つの想定可能な結果が、図13に示す特徴集合1300によって例示されている。図13に示された6成分特徴集合1300は、図6に示した係数位置610～615を含む。図13に示された6成分特徴ベクトル1300の係数位置1301として含まれ、図6に示された変換マトリックス600の第2行、第6列の係数位置614の包含は、11f<sub>11</sub>に対応する比較的高い水平頻度成分が画像クラスを弁別する際に有効であることを示す。高頻度成分の包含はほとんど、フレームを認識するのに一般に比較的小さく鋭いエッジを有するテキスト等の小さな鋭い特徴を検出することを要する場合に生じる。

【0037】図14は、本発明に従って、図13に示す6成分特徴集合を有するトレーニングフレームの2個以上の特徴ベクトルから計算された平均特徴ベクトル1400を示す。

【0038】図15は、本発明に従って、図13に示す特徴集合を有するトレーニングフレームの2個以上の特徴ベクトルから計算された対角共分散マトリックス1500を示す。特徴集合で示された係数位置の値の間に相関関係が存在しないという仮定にもとづき、対角共分散マトリックス1500の非対角線上成分は、やはりゼロに設定されている。

【0039】図16は、本発明に従った分類のための図13に示す特徴集合1300を有するフレームから検索された特徴ベクトル1600を示す。このように、成分1601～1606は、本発明の方法に従って分類されるフレームの変換によって生じる変換マトリックスから得られる実際の個々の変換係数を表現している。

【0040】十分なデータ簡約化が行われた場合、クラシファイヤは、プレゼンテーションスライド、発表者または聴衆といった典型的な会議のビデオシーンの間を識別するために、本発明に従って容易にトレーニングされる。会議ビデオの領域分野の他に、この方法は、ニュースキャスターのショットなど、あるクラスの画像が類似の構成を有する場合には良好に作用するはずである。本発明による方法を評価するために、ビデオテープ録画されたスタッフ会議の資料に関して多数の実験を行った。ビデオショットは、6つのカテゴリに分類され、資料はトレーニング集合および試験集合に分けられた。

【0041】ビデオ分類実験は、6カ月の期間に開かれビデオ録画されたスタッフ会議の資料に関して実施された。各ビデオはカメラオペレータによって作成され、オペレータは、パン/チルト/ズームの制御が可能な3台のカメラによるビデオと、パーソナルコンピュータおよび演壇カメラからのビデオ信号との間で切り換えを行った。後者の装置は、透明および不透明資料といったプレゼンテーショングラフィックスを背面映写スクリーンに表示できるようにした。従って、ビデオショットは一般に、発表者、聴衆ショットおよび、「パワーポイント

（登録商標）Ｊスライドまたは透明資料といったプレゼンテーショングラフィックスより構成される。得られたビデオはMPEG-1符号化され、サーバに記憶された。

【0042】資料には、21回の会議ビデオが収められ、ビデオ収録時間の合計は13時間以上であった。資料は、会議ビデオを交互に選ぶ形で、試験およびトレーニングセグメントに任意にセグメント化された。試験およびトレーニングデータは下記の表1に示す6つのクラスにラベルづけられ、表はトレーニングおよび試験の各集合のフレームの数も示している。相当量のデータがいずれのカテゴリにも当てはまらず、ラベルづけされずに残された。6つのクラスは、プレゼンテーショングラフィックス（slides）、照明付き映写スクリーンのロングショット（longsw）、照明なしの映写スクリーンのロングショット（longsb）、聴衆のロン\*

ショットカテゴリ	トレーニングデータ	試験データ
slides	16,113	12,969
longsw	9,102	5,273
longsb	6,183	5,208
crowd	3,488	1,806
figonw	3,894	1,806
figonb	5,754	1,003
カテゴリなし	13,287	10,947
合計	67,821	39,047

【0044】実験は、ガウス分布型クラシファイヤが長時間のビデオの脈絡における特定のクラスからビデオフレームを検出することを実証している。これは、長時間のビデオから、類似フレームの領域として定義されたショットをセグメント化するために使用される。これは、例えばスライドを含むショットの始まり、といった有益な索引点を提供する。他の方面では、例えばフレームまたは色の相違によってショットがすでに突き止められている場合、そのショットから全部のフレームに関してショットモデルが容易にトレーニングできる。これにより、共分散が動きその他の変化により生じる差異を捕捉するので、ショットを類似性により検索できるようになる。あるショットを表現するキーフレームが、尤度距離計量を用いてそのショット平均に最も近いフレームを見つけることにより容易に見出せる。画像を表現する係数の数は極めて控え目であるので（主成分分析のフレーム当たり10個程度の特徴）、1つの代替法は、ビデオデータ自体に比べてもほとんどまったくオーバーヘッドを伴うことなく、ビデオとともに特徴を記憶することである。ガウスモデルは計算が容易なので、モデルは選択的にオンザフライでトレーニングされる。これは対話型ビ

\*グショット（crowd）、明背景での人物像の中間クローズアップ（figonw）および暗背景での人物像の中間クローズアップ（figonb）を表現するように選択された。（スクリーンショットといった）単一のカテゴリでかつ（照明付きと照明なしのスクリーンショットといった）著しく異なるモードの場合、各モードについて別のモデルが使用された。これは、単一ガウスモデルとのすぐれた一致を保証したが、別の方法は結合モデルをモデル化するためにガウス混合を代替的に使用する。同一の論理クラスをモデル化するように意図されている場合、異なるモデルは選択的に結合される。例えば、人物像を見つけることを意図している場合に背景色は重要ではないので、分類結果を提示する際にfigonwおよびfigonbクラスの結合が行われる。

【0043】

【表1】

デオ検索のような用途を可能にし、この場合、ユーザは、時間バー上をドラッグしてビデオ領域を選択するなどして、所望のクラスを指示する。モデルはその領域の特徴について迅速にトレーニングされ、大きなビデオ資料のフレームに対応する類似性が迅速に計算される。資料における高い尤度の領域は、選択されたビデオに良好に一致する領域であり、資料の索引として機能する。

【0045】多様なモデル結果をスレッシュホールド設定を用いずに示すために、最大尤度法を使用して、ラベルづけされた試験フレームを分類した。下記の表2は、30個の最大分散離散コサイン変換係数の使用による結果を示す。クラスfigは、figonwおよびfigonbの結合クラスの上位集合である。各列は試験フレームの实地検証情報ラベルであり、行は、行クラスとして認識される試験集合におけるサンプルの割合（小数分率）を示す。非ゼロの非対角線上成分は分類誤差を表す。すべてのラベルづけされたフレームは、それ自体のラベルと異なることはあっても最大尤度クラスを有するので、列の合計は1になる。

【0046】

【表2】

	slides	longsw	longsb	crowd	fig
slides	0.872	0.017	0.000	0.000	0.000
longsw	0.009	0.900	0.000	0.000	0.000
longsb	0.000	0.002	0.749	0.000	0.000
crowd	0.001	0.042	0.014	0.848	0.010
fig	0.118	0.039	0.237	0.152	0.990

【0047】図17は、本発明によって類似性を決定する方法において、類似性を決定するためのスレッシュホールドとして使用されるスライド画像クラス統計モデルの標準偏差の倍数の関数として、スライドとして正しく識別されたスライドフレームの割合およびスライドとして誤って識別された非スライドフレームの割合を示す。代替的な実施の形態として、類似性を決定するためのスレッシュホールドは一般的なものであり、例えば、他のクラスの最大尤度によって決定される。x軸は標準偏差の所定の倍数を表し、y軸はスレッシュホールドのその特定の選択にもとづいて類似と識別されたフレームの割合を表す。プロット1701は、実際にスライドであり、本発明の類似性評価法によって正しくスライドとして識別されたフレームの割合を示す。プロット1702は、実際にはスライドでなく、本発明の類似性評価法によって誤ってスライドとして分類されたフレームの割合を示す。

【0048】図17は、ビデオを分類しセグメント化するためにガウスモデルがどのように使用されるかを実証している。スタッフ会議ビデオの資料による実験は、スライド、話者、聴衆といったクラスが正しく認識されることを示した。1/2秒間隔で取られたMPEG-1フレームは、復号化され、64×64グレイスケール強度下位画像に簡約化された。得られたフレーム画像は離散コサイン変換およびアダマール変換により符号化され、最大平均値を有する100個の係数が特徴として選択された。対角共分散ガウスモデルが80例のスライドフレームについてトレーニングされ、無関係の試験ビデオのスライドフレームおよびタイトルの確率を計算するために使用された。

【0049】(共分散 $\Sigma^{-1/2}$ からの)標準偏差の倍数で尤度をスレッシュホールドとすることは、クラスの帰属関係を検出するうえで極めて有効であることが示されている。また、そのようなスレッシュホールドは、使用される係数の数からまったく独立である。図17は、スライド検出比が多様なスレッシュホールドにおいてどのように変化するかを示している。グラフは、スレッシュホールドが標準偏差の約1.1倍のとき、84%の正確なスライド認識率をもたらし、誤り認識はほとんどない(9%)ことを示している。標準偏差により正規化された場合、尤度は、それ自体で、クラスモデルに対する所与のフレームの類似性の指標として有益である。全部のクラスが同様の検出率を有するが、誤り認識の数はクラスごとに異なっている。

【0050】単純ガウスモデルは上記のようにトレーニング画像の平均を計算するので、画像シーケンスに関係するいずれかの時間変化情報を失う。動きまたは連続といった動的な情報を捕捉するために、モデルは様々な方式で選択的に拡張される。フレーム間差異または簡約化された特徴の傾向によりモデルをトレーニングすることにより、動きまたはフェードといった時間変化効果はモ

デル化される。

【0051】図18は、本発明による画像クラス統計モデルを用いたビデオフレームの類似性を判定する方法を示す。工程1801で、現在分析中のフレームに対応する特徴ベクトルが検索される。工程1802で、その画像クラス統計モデルに対応する平均特徴ベクトルが検索される。工程1803で、特徴ベクトルからの平均特徴ベクトルの減算を表現する差分ベクトルが計算される。工程1804で、差分ベクトルの大きさが画像クラス統計モデルの標準偏差の所定の倍数と比較される。差の大きさが標準偏差の所定の倍数より小さい場合、工程1805はそのフレームを類似として分類する。差の大きさが標準偏差の所定の倍数より小さくなければ、工程1806がそのフレームを非類似として分類する。図18に例示した類似性を判定する方法は、ガウスの公式による実際の確率計算を必要としないことに留意しなければならない。代わりに、差分ベクトルの大きさおよび標準偏差がユークリッド距離として計算される。差分ベクトルの大きさは、そのd個の成分の平方の和の平方根によって計算される。画像クラスの標準偏差は、対角共分散マトリックスの対角線上成分の和の平方根として計算される。

【0052】図19は、本発明に従ってビデオの各種フレームを生成する画像クラス統計モデルの確率の対数表示を示している。対数は単調関数なので、確率の対数は、類似性の大小を判定するために確率を比較するのと同様に比較される。

【0053】図19は、ほぼ1時間継続する試験ビデオにおけるスライド画像でトレーニングされたガウスモデルの対数尤度を示す。ビデオでスライドが実際に示された時を指示する「実地検証情報(ground truth)」は、上部付近の幅広のバーとして示されている。この対数尤度が、ビデオでスライドが示されている時の良好な指標であることは明らかである。(共分散マトリックス $\Sigma_c$ から計算された)ある標準偏差で対数尤度をスレッシュホールドとすることは、個々のフレームを分類するうえで極めて効果的であることが示されている。

(共分散から計算された)標準偏差の倍数で尤度をスレッシュホールドとすることは、クラスの帰属関係を検出するうえで極めて効果的であることが示されている。また、こうしたスレッシュホールドは使用される係数の数とはまったく独立である。

【0054】いずれかの特定のフレームまたはフレームのビデオセグメントと画像クラスとの間の類似性は、本発明に従って計算される。ガウスモデルの場合、所与のフレームの類似性測度は尤度であり、対数領域にあるものに代えることができる。ガウスモデルはまた、セグメント境界として機能する所与のスレッシュホールドを類似性測度が超えた時にそれらのフレームを見つけることによって、ビデオをセグメント化するためにも使用できる。

時間モデルが存在しない場合、最小のセグメント長を求めるというような臨時の規則によりセグメント化を改善できる。

【0055】図20は、本発明に従ってビデオの各種フレームを生成するビデオ画像クラス統計モデルの確率の対数を表示する方法を示す。工程2001で、画像クラス統計モデルによって生成されるフレームの確率がガウスの公式によって計算される。工程2002で、確率の対数が計算される。工程2003で、確率の対数が図19と同様にして表示される。工程2004において、さらにフレームが存在する場合は、2006に分岐して工程2001に戻り、それ以上フレームがない場合、工程2005で終了する。

【0056】図21は、特徴集合の成分の数dの関数として正しく分類されたフレームの割合、それらのフレームに適用された変換のタイプおよびd成分特徴集合の選択方法を示している。図21は、離散コサイン変換およびアダマル変換の両者について、正しい分類の確度が、一般に、特徴集合が増加するにつれ変換係数の数とともに向上することを示している。トレース2101、2102および2103の下降部分は、各クラスがそうした大きな数の係数位置を有する特徴集合を判定するために十分なトレーニングフレームがトレーニング集合に存在しないことの結果である。言い換えれば、トレース2101、2102および2103の下降部分は、特徴ベクトルが由来するべきものとして合理的にモデル化されるガウス分布に倣うのではなく、トレーニングフレームの特徴ベクトルの実際のデータポイントに倣っていることを示している。分布に倣わせるためには、トレーニングフレームの数は特徴集合の変換係数の数よりも相当に多くなければならない。これは、与えられた数のトレーニングフレームを前提とする限り、特徴集合の変換係数位置を100以下とすることが、計算上の負荷を軽くするだけでなく、より大きな特徴集合よりも効果的であることを実証する。

【0057】異なる変換方法での変換係数の数の影響を判定するために、全正確さ、すなわち、正しいカテゴリに認識されたサンプルの割合を計算した。図21はその結果を示す。離散コサイン変換およびアダマル変換の主成分に関する認識分布がほぼ同一であることを指摘できるのは興味深い。最良の成績(87%正確な)は10個の主成分を用いて得られた。主成分分析を伴わない場合、分散順位づけ離散コサイン変換係数は30をピークとするのに対し、アダマル変換は300で若干高い確度を得る。アダマル変換はしばしば、離散コサイン変換と同様に知覚的特徴を保存しないということで批判されるが、この場合には多少すぐれているように思われる。直線アダマル変換ベースの関数は、シヌソイド離散コサイン変換系よりも良好に(スライドや壁といった)画像特徴を一致させるからである。

【0058】図22は、本発明の方法に従ってスライドに類似であると見られるビデオの領域を表示するブラウザを示している。ブラウザ2200は、スライドビデオ画像クラスに類似であると判断されるフレームより構成されるビデオ内の時間間隔を黒い垂直バーで示す時間バー2201を含む。

【0059】ユーザがビデオ内の興味のある部分を見つけるのを助成するためにビデオ分類を使用するアプリケーションが本発明に従って開発されている。長時間のビデオがそれを全体として見ることなく所望の情報を含むかどうかを判定することは単純ではない。インテリジェントメディアブラウザは、図22に示すように、ビデオから抽出されたメタデータを利用することによってビデオに対するきめ細かいアクセスを可能にする。あるビデオに関する信頼度スコアが時間バーにグラフィカルに表示される。信頼度スコアは、ソースメディアストリームへのランダムアクセスに時間軸を使用することによりソースストリームにおける興味ある領域への貴重な糸口を付与する。例えば、スライドモデルの正規化対数尤度が図22の時間バーに表示される。高尤度(信頼度)の2つの領域が灰色または黒色領域として視覚化され、それらはビデオにおけるスライド画像に対応する。時間軸上の点または領域を選択すると、対応する時間からメディアの再生を開始する。このようにして、興味のある部分となる高い可能性の時間間隔が、信頼度表示から視覚的に識別され、線形探索を伴わずに容易に調査できる。

【0060】図23は、本発明に従ってビデオを分類する方法において使用される隠れマルコフモデルに対応するクラス遷移図を示す。画像クラスG、AおよびBの各々はガウス分布を用いてモデル化される。同一のクラスに留まるかまたは別のクラスに遷移する遷移確率は、遷移矢印の横に示されている。

【0061】隠れマルコフモデルは、本発明に従ってビデオセグメントの継続時間およびシーケンス(順序)を明示的にモデル化できる。単純な実施例では、2状態隠れマルコフモデルの一方の状態は所望のクラスをモデル化し、他方の状態モデルは他のすべてをモデル化する

(「ガーベージ」モデル)。多状態隠れマルコフモデルは、上記のガウスモデルを用いて、それらを平行に結合し、弧に沿って遷移ペナルティを加えることによって作成される。図23は、そうしたモデルを示しており、状態Gがガーベージモデルであり、状態AおよびBが所要のビデオクラスをモデル化している。(図示されたシーケンスは、ビデオクラスが2つの個別の成分AおよびBを有し、AがBの前に生起することを示唆している。多数の他のモデルシーケンスが可能である。)ビデオに対する最大尤度を使用した隠れマルコフモデルのアライメントはバイタービアルゴリズムによって決定される。これは、サンプルと類似しているセグメントおよび類似でないセグメントへのビデオのセグメント化をもたらす。

さらに、観測されたビデオを生じるいずれかの特定の状態の尤度は、いずれかの特定のフレームについて選択的に決定され、探索、順位づけまたはブラウジングにおいて活用するための有用な類似性測度を与える。

【0062】図23は、尤度スレッシュホールドを有する単一のガウスモデルが長時間のビデオから類似のショットをどのようにしてセグメント化できるかを示している。多様なショットモデルを使用することにより、尤度比または最大尤度を用いて、いずれのモデルにも良好に一致しないショットを排除するスレッシュホールドを選択的に有する多様なショットをセグメント化できる。異なるショットは、多様な代替計量を用いてそれらのガウスモデルを比較することによって、比較照合される。

【0063】クエリー状態の隠れマルコフモデル出力分布は、上記のガウスモデルに関してまさに説明した通り、係数特徴の単数または複数のガウスモデルとして代替的にモデル化される。選択的に、エルゴード的に（完全に）結合された複数の状態が、複数の混合ガウスモデルと同様にセグメントをモデル化するために使用される。単数または複数のガーベージモデルの出力分布もガウス分布である。そのパラメータは、ビデオデータベースから推定され、システムに記憶される。クエリーおよびガーベージ状態に留まる遷移確率は、例題データから推定されるかまたは、クエリーの長さおよびビデオにおけるクエリーの生起間の長さが変化し得るので、ユーザによって選択的に調整される。この方式の利点は、遷移確率がほとんどの隣接フレームを同一状態に拘束し、従って見かけ上のセグメント化または類似性スコアの変動を低減することである。

【0064】隠れマルコフモデルの公式化は、複数の状態および（音声認識における言語モデルに類似の）遷移グラフを用いてビデオの傾向またはシーケンスを捕捉するために強力に拡張されている。それ故、隠れマルコフモデルは、例えば、ニュース放送の開始を特徴づける放送局のロゴからニュースキャスターのショットへの遷移をモデル化するために選択的に使用される。この例で図23について説明すれば、状態Aは放送局のロゴをモデル化し、状態Bはニュースキャスターのショットをモデル化する。隠れマルコフモデルにおける暗示的シーケンス拘束のために、これは、A-Bシーケンスにのみ一致しB-Aシーケンスには一致せず、または、AまたはBを孤立してセグメント化するのに対して、単純ガウスモデルは全部について高いスコアを生じる。

【0065】図24は、図23に示したクラス遷移図に対応する本発明に従ったクラス遷移確率マトリックスを示している。クラス遷移確率マトリックス2400の行は以前のフレームのクラスを表し、マトリックス2400の列は現在フレームのクラスを表す。クラス遷移確率マトリックス2400の各列は、ある現在の画像クラスに関係するクラス遷移確率ベクトルである。図23に示

したクラス遷移図は以降のフレームについてクラスGからクラスBへの遷移を許していないので、マトリックス2400の成分2401はゼロである。同様に、クラス遷移図2300はクラスBからクラスAへの遷移を許していないので、マトリックス2400の成分2402はゼロである。

【0066】図25は、図23に示したクラス遷移図に従った5つの連続した初期ビデオフレームに対応する可能なクラスシーケンスの全てを示す。クラス遷移図2300はそのシーケンスがクラスGで始まるように指示しているので、最初のフレームのクラスは図25の枠2501に示されたGである。しかし、第2のフレームは、それぞれ枠2502および2503に示されたクラスGまたはクラスAのどちらか一方となる。第2のフレームが枠2503で示されたクラスAである場合、第3のフレームは、それぞれ枠2504、2405および2506に示されたクラスG、AまたはBのいずれかとなる。クラスの確率は、そのクラスについて計算された尤度、以前のクラスの確率および、そのクラスへの遷移を生じるクラス遷移確率の関数である。各状態の確率は以下の式によって与えられる。

【0067】

【数2】

$$P_G(t) = P_G(X) \cdot \max \left\{ \begin{array}{l} P_G(t-1) \cdot P_{GG} \\ P_A(t-1) \cdot P_{AG} \\ P_B(t-1) \cdot P_{BG} \end{array} \right\} \quad \text{式1}$$

$$\text{if } G(t) \text{ and max is } \left\{ \begin{array}{l} P_G(t-1) \cdot P_{GG} \text{ then } G(t-1) \\ P_A(t-1) \cdot P_{AG} \text{ then } A(t-1) \\ P_B(t-1) \cdot P_{BG} \text{ then } B(t-1) \end{array} \right. \quad \text{式2}$$

$$P_A(t) = P_A(X) \cdot \max \left\{ \begin{array}{l} P_G(t-1) \cdot P_{GA} \\ P_A(t-1) \cdot P_{AA} \end{array} \right\} \quad \text{式3}$$

$$\text{if } A(t) \text{ and max is } \left\{ \begin{array}{l} P_G(t-1) \cdot P_{GA} \text{ then } G(t-1) \\ P_A(t-1) \cdot P_{AA} \text{ then } A(t-1) \end{array} \right. \quad \text{式4}$$

$$P_B(t) = P_B(X) \cdot \max \left\{ \begin{array}{l} P_A(t-1) \cdot P_{AB} \\ P_B(t-1) \cdot P_{BB} \end{array} \right\} \quad \text{式5}$$

$$\text{if } B(t) \text{ and max is } \left\{ \begin{array}{l} P_A(t-1) \cdot P_{AB} \text{ then } A(t-1) \\ P_B(t-1) \cdot P_{BB} \text{ then } B(t-1) \end{array} \right. \quad \text{式6}$$

【0068】図26は、本発明によるクラス遷移確率マトリックスおよび画像クラス統計モデルによってビデオをセグメント化する方法を示している。方法は工程2601に始まる。工程2602で、可能性のある現在の状態の各々に対応する最も確からしい以前の状態が計算される。それらの計算は図25に示した例に関する上記の式を用いて行われる。工程2603で、現在のフレームの尤度が、各画像クラスに対応するガウス関数によって

可能な現在の状態の各々について計算される。工程2603での計算は、例えば図12に示した方法1200の工程1204において、計算された確率と同一である。工程2604で、全部の可能な状態に対応する現在の状態の確率が工程2603および2602による結果を用いて計算される。工程2604の計算は上記の式によって実行される。工程2602の計算は、現在の状態を仮定して式2、4および6を使用する。工程2604の計算は、上記の式1、3および5を使用する。検査2605はビデオの終わりに達したかどうかを判断し、否定されれば、工程2606はプロセスを次のフレームに進める。それが最後のフレームであれば、工程2605は処理を工程2606に渡し、そこでその最終状態が最大の全確率を有する状態として選択される。最終状態が選択された後、最も確からしい以前の状態が、上記の式2、4および6の以前の評価に従って選択される。言い換えれば、最終状態が既知であれば、以前の状態の全ては、工程2602ですで行われた計算によって自明になる。工程2608で、さらにフレームが存在するかどうか判定され、肯定されれば、工程2609はその以前のフレームを工程2607に渡し、工程2602ですすでに計算された結果に従って次の以前の状態とのリンクの決定がなされる。第1のフレームが分類されると、処理は工程2610で終了する。

【0069】隠れマルコフモデルの場合、セグメント化は、最大尤度状態シーケンスを見つけるためのバイタービアルゴリズムによって行われる。これは、特定の状態または状態の群とアライメントされた全部のフレームがセグメントとしてみなされるので、最大尤度セグメント化を直接与える。隠れマルコフモデルの構造は、アライメントが（従来行われていたように局所的にではなく）ビデオ全体について計算されるので、このタスクに特に適している。このモデルに内在するシーケンスおよび継続時間の拘束は、他の方式の分類誤りによって生じ得る単一フレームセグメントといった誤りを効果的に禁止する。所与のフレームとクエリーとの間の類似性は、バイタービアルゴリズムにおいて、単数または複数のクエリー状態の事後確率として計算される。類似性測度が与えられ、ビデオのあらゆる集合は、クエリーセグメントとの類似性によってセグメント化および／または順位づけられる。これは、ビデオの大きな資料からの類似性による内容にもとづく検索を可能にする。

【0070】上述のように単純ガウスモデルはトレーニングフレームの平均を計算するので、ビデオシーケンスに関係する何らかの時間変化情報を失う。動的なシーケンス情報を捕捉するために、モデルは多様な方法で選択的に拡張される。フレーム間の差異または簡約化された特徴の傾向でモデルをトレーニングすることによって、動きまたはファクシミリといった時間変化効果はモデル化される。ビデオシーケンス間の類似性を見つけるため

に、2つのシーケンスのフレームのフレームごとの内積を合算することにより相関スコアが計算される。類似なシーケンスは大きな相関を有する。異なる長さの2つのシーケンス間の最良の一致を見つけるために動的プログラミングが選択的に使用される。本発明による動的な事象を捕捉するすぐれた技法は、特徴出力確率をモデル化するためにガウス混合を用い、特に音声認識用に開発された効率的なトレーニングおよび認識アルゴリズムが与えられた、隠れマルコフモデルである。

【0071】ここで行った実験は、変換係数の統計モデルが低い誤差率でビデオフレームを迅速に分類することを実証している。この方式の計算の単純さおよび少ない記憶要求量は、本発明による対話型ビデオ検索といった用途を可能にする。

【0072】特定のビデオセグメントについてビデオデータベースを探索する際に、所望のビデオセグメントのタイプの記述を与えるよりも、例題を与えることによってクエリーを指定するほうが容易であることが多い。例えば、話を聞いている一群の人々を示すビデオのセグメントが望まれる場合、探索クエリーとしてシステムに群衆セグメントを単純に呈示することはより容易である。これは、選択されたセグメントに類似であるセグメントについて単一のビデオを探索する際に特に当てはまる。類似性による検索は、ユーザにとって容易であることに加え、実例からクエリーの良好なモデルを作成することが容易であるので、より正確であることが多い。

【0073】自動ビデオ分類は、ブラウジング、自動セグメント化および内容にもとづく検索といった広範な用途に有用である。自動分類を用いたアプリケーションは、特定の話者を示すビデオを検索するか、または、ビデオの再生中にその話者のいる領域を強調表示させるなどによって、デジタル化ビデオをブラウジングおよび検索するうえでユーザを支援することができる。自動生成注釈は、ビデオテープ録画された会議から重要な情報を検索する際にユーザを支援することができる。このようなツールは、ユーザが、特定のビデオおよびそのビデオ内の対象となる領域の両方を突き止めなければならない場合に、ビデオの大きな集合を取り扱うのを助けることができる。こうしたあらゆる用途にとって、ビデオのトレーニング用集合は異なるビデオおよびオーディオクラスに従ってラベルづけされ、統計モデルはそのラベルづけされたセグメントでトレーニングされる。

【0074】本発明は、ビデオの類似性の統計的測度および、その類似性測度を使用して再生中にビデオの案内を助成するアプリケーションを含む。本発明によれば、類似性マッチングに使用されるビデオの領域を選択するための2つの異なるユーザインタフェースが開示される。

【0075】ブラウザは、ビデオ領域を選択し類似領域を自動的に見つけることによってユーザにビデオの構造

を探索させるように設計されている。例えば、ニュース放送を見る場合、ユーザはニュースキャスターのショットを含む領域を選択する。システムはその後、類似の領域を自動的に検出し、それらをグラフィカルに表示しつつ自動索引点として示し、それによりユーザは、例えば、介在箇所を見ることなく次の類似領域に直接跳ぶことが可能になる。これらの索引は、以後のユーザのために保存し注釈を付けることができる。類似性索引は対話的にかつ極めて迅速に作成できる。

【0076】図27は、本発明に従って類似性探索を実行する方法におけるデータの流れを示している。ソースビデオ2701は、トレーニングセグメントが抽出されるビデオを表す。変換特徴2702は、図2において変換特徴208がビデオファイル201から抽出されたのと同様にして、ソースビデオ2701から抽出される。工程2703は、トレーニングフレームの収集のためのトレーニング領域のユーザ選択を示している。工程2704で、ガウス画像クラス統計モデルが、平均特徴ベクトルおよび対角共分散マトリックスを比較することによりトレーニングされる。ビデオ2705は、類似性の探索のためのターゲットとされたビデオを表す。同様に、変換特徴2706が抽出される。工程2707において尤度計算が、工程2704でトレーニングされた画像クラス統計モデルを用いて行われ、得られた確率が工程2708でフレームごとに出力される。

【0077】図27は、システムが実際にどのように使用されるかのブロック図を示す。ユーザは最初に単数または複数のビデオセグメントを選択することによりクエリを実行する。クエリの簡約化された離散コサイン変換またはアダマール変換係数が、オンザフライでの計算またはデータベースにルックアップのどちらか一方によって得られる。クエリのモデルはその後これらの係数を用いてトレーニングされる。単純な場合、単純ガウスモデルが使用される。データベース内のビデオの簡約化された離散コサイン変換またはアダマール変換係数はシステムに提示され、尤度計算が実行される。これは、一連の類似性スコアおよび、類似および非類似セグメントへのセグメント化を生じる。類似性スコアはその後ブラウザに表示され、ユーザが類似のビデオセグメントを調査できるようにする。

【0078】類似性計算のデータは、図2の説明において前述したものと同様にして離散コサイン変換またはアダマール変換のどちらか一方によって得られる。この表現は、類似画像のフレームが類似の特徴を有するので、類似性を測定するために適切である。

【0079】変換法にもとづく類似性測度は、従来のカラーヒストグラム方式よりも多くの用途に関してすぐれている。特に、変換係数は、形状についてほとんど変化がないヒストグラムと異なり、画像における主要な形状およびテクスチャを表現する。例えば、左上および右下

に同一物体がある2つの画像は、ヒストグラムでの相違はごくわずかであるが、本発明による変換ドメインにおいては顕著に異なる。現在の類似性測度は輝度だけにもとづいているが、後述の通り、この技法を色を使用するように拡張することは容易なはずである。

【0080】この変換法により可能なセグメント化およびモデル化の種類が比較的粗いことを指摘することは重要である。例えば、ニュース放送においてニュースキャスターとロケーションのショットとを識別することは単純であるが、特定のニュースキャスターを識別するといった、より精緻な区別はさらに特殊化されたデータ簡約化またはドメイン特定モデルを必要とするであろう。しかし、これらの技法は、例えば、群衆または自然のシーンを排除しつつ計算上高価な顔面識別アルゴリズムにより、以後の分析のために適切なクローズアップシーンを選択するといった、より精巧な方法の重要なフロントエンドまたはプレクラシファイヤとして代替的に機能する。

【0081】図28は、本発明に従ってビデオに対応する特徴ベクトルデータベースを計算する方法を示している。迅速な尤度計算および画像クラス統計モデルの迅速なトレーニングを助成するために、ビデオのフレームに対応する特徴ベクトルを予備計算し、それを特徴データベースに記憶することが望ましい。工程2801で、フレームが離散コサイン変換またはアダマール変換によって変換される。工程2802で、変換係数マトリックスから特徴ベクトルが抽出される。工程2803で、特徴ベクトルが特徴ベクトルデータベースに記憶される。検査2804では、さらにフレームがあれば、次のフレームが工程2801に渡され、それ以上フレームがなければ、方法は工程2805で終了する。

【0082】ビデオ領域間の類似性を評価するために、ビデオフレームの類似性が開示される。各フレームは、離散コサイン変換またはアダマール変換といった正規直交射影によって変換される。変換が、下位ブロックではなく画像全体について行われた場合、係数は画像を正確に表現する。変換されたデータはその後、上述のように切り捨て、主成分分析または線形識別解析などのいずれかの技法によって簡約化される。ここに提示した用途の場合、最大分散係数以外の全部を破棄することが良好に作用する。その簡約化表現は、高度にコンパクトであり、元のフレームの顕著な情報を保存している。これは、元の画像を復元することを意図する、データ圧縮とは異なることに留意されたい。元のデータは表示および使用に利用可能であると前提されているので、変換プロセスを逆にする必要はまったくない。従って、この変換法は、コンパクト性または画像忠実度よりも分析のために最適化されている。

【0083】結果として得られるのは、各フレームのコンパクトな特徴ベクトルまたは簡約化された係数（10

～30パラメータ)である。この表現は、類似のフレームは類似の変換係数を有するので、ビデオの類似性を数値化するために適切である。特定のショットと隣接するフレームといった類似画像の集合をモデル化するために、ガウスモデルが例題フレームでトレーニングされる。ガウスの平均は例題フレームの平均を捕捉し、共分散は動きまたは照明の相違による変動をモデル化する。単一混合ガウスは、例題データに関して1パスで極めて迅速に選択的に計算され、例題フレームのおおよその構成および可変性をモデル化する。

【0084】多くの用途にとって、完全なビデオフレームレートは必要なく、フレームは、毎秒数フレームだけを変換する必要があるような時間で間引かれる。こうした要因は、記憶コストが実際上無視でき、係数が計算されれば計算時間は極めて迅速であることを意味する。従って、リアルタイムアプリケーションに使用される戦略は、簡約化された係数を予備計算し、それらをビデオとともに記憶し、対話的かつ迅速な類似性測定を可能にすることである。MPEG-7といった将来のフォーマットはそうしたメタデータをビデオデータとともに含めることを可能にするが、現在好ましい実施の形態による用途では、係数は個別のファイルに記憶される。

【0085】図29は、本発明に従って統計モデルを対話的にトレーニングする方法を示す。工程2901で、トレーニングフレームまたはトレーニングセグメントがユーザにより対話的に選択される。工程2902で、工程2901で選択されたトレーニングフレームまたはセグメントに対応する特徴ベクトルが、直接の計算または特徴ベクトルデータベースのルックアップのどちらか一方によって得られる。工程2903で、トレーニングフ

レームに対応する特徴ベクトルから平均特徴ベクトルおよび対角共分散マトリックスを計算することによって、画像クラス統計モデルが構築される。

【0086】変換ドメインの1つの利点は、フレームを表現する特徴ベクトルの大きさが極めて控え目である(PCA特徴についてフレーム当たり10程度)ということである。クエリービデオトレーニングセグメントは、平均ベクトルおよび共分散マトリックスによってパラメータ化された多次元ガウス分布によりモデル化される。実際、特徴間のゼロ相関が前提とされるように対角共分散マトリックスを仮定することは普通であり、各特徴はガウス分布を有する独立のランダム変数であると仮定される。対角共分散マトリックス(すなわち非対角線上の成分がゼロである)は、モデルが高次元で頑強性を持つ(ロバスト)であるように仮定されている。ガウスモデルを用いてクラスをモデル化するために、トレーニング画像の集合について平均および共分散が計算される。クエリートレーニングセグメントは、平均ベクトルおよび共分散マトリックスを計算するために使用される。類似性スコアは、ビデオの各フレームについて、ク

エリー画像クラス統計モデルからフレームの尤度を計算することによって計算される。代替的に、より精巧なモデルは、ガウス混合を使用し、期待値最大化アルゴリズムを利用して、複数のパラメータおよび混合重み、それにより、複数のガウスモデルの各々に関する複数の平均、分散および重み係数を評価する。しかしこれは、反復を要する。そうしたわけで、オンザフライで迅速に計算される単一混合ガウスモデルが仮定されている。

【0087】フレームの係数に平均値を設定し、分散を定数等の値に設定することによって、またはいずれかのトレーニング集合から得られた分散を使用することによって、ガウスモデルを生成するために単一フレームクエリーが選択的に使用されることに留意されたい。他のフレームまたは静止画像はその後、類似性についてスコアが付けられる。定数の分散はユークリッド距離計量を生じ、トレーニング分散はマハロノビシュ(mahalanobis)距離を生じる。従って、類似の静止フレームまたは画像は、それらを距離測定によって順位づけることによって集合から検索される。本発明によるこのシステムの別の変種は、ただ1個の画像をクエリーとして使用する従来の画像検索システムではなく、画像の群またはクラスでクエリーモデルがトレーニングされた場合である。

【0088】一度計算されると、任意のビデオフレームの類似性は、モデルがフレームを生成する尤度によって決定される。類似フレームは高い尤度を生じる。この方式は、会議ビデオの大きな資料での話者およびスライドといった所定のビデオクラスについて約90%の分類率をもたらしている。ガウスモデルは、動きまたは照明の相違による変動をモデル化しつつ、画像クラスの特徴的な構成および形状を捕捉することができる。特徴ベクトルが計算されると、多数の用途が使用可能である。最も単純なものの1つは直接的な距離測定である。類似フレームは類似の特徴ベクトルを生じるので、特徴ベクトル間の距離を測定することにより画像距離の指標が得られる。

【0089】図30は、本発明に従ってブラウザ内にビデオフレームを呈示し、類似性測定を表示する方法を示す。工程3001でフレームの特徴ベクトルが検索される。工程3002で、画像クラス統計モデルによって生成される特徴ベクトルの確率が計算される。工程3003で、その確率がスレッシュホールドより大きいかが判定される。スレッシュホールドはやはりユーザによって対話的に定義される。工程3002で計算された尤度がスレッシュホールドより大きければ、工程3004はそのフレームを類似として索引づける。尤度がスレッシュホールドより小さければ、そのフレームを工程3005で非類似として索引づける。工程3006で、類似または非類似の類似性属性はそのフレームについてブラウザにグラフィカルに表示される。

【0090】いずれかの特定のフレームまたはビデオセグメントとクエリーセグメントとの間の類似性が計算される。ガウスモデルの場合、所与のフレームの類似性は尤度であり、代替的に対数ドメインに存在する。ガウスモデルはまた、セグメント境界として機能する、また、所与のスレッシュホールドを類似性が超えた場合に、それらのフレームを見つけることによってビデオをセグメント化するためにも使用される。継続時間モデルが存在しない場合、最小セグメント長を要求するような臨時的規則がセグメント化を改善させることができる。

【0091】図31は、本発明に従って、対話的に定義されたトレーニングビデオセグメント、そのトレーニングビデオセグメントのトレーニングフレームから得られた平均特徴ベクトルの逆離散コサイン変換、およびトレーニングビデオセグメントのトレーニングフレームから得られた平均特徴ベクトルの逆アダマール変換を示す。フレーム3101はユーザによって対話的に定義されたトレーニング画像を表す。フレーム3102は、フレーム3101に示すトレーニング画像から得られた平均特徴ベクトルの逆離散コサイン変換を表す。フレーム3103は、フレーム3101に示すトレーニング画像から得られた平均特徴ベクトルに対応する逆アダマール変換を表す。

【0092】ビデオ類似の領域を突き止める本発明に従った方法は既述の通りである。類似性測度を用いるビデオブラウザを提供する、直接的なアプリケーションを以下に述べる。図32は、1つのブラウザのプロトタイプユーザインタフェースを示す。左上に通常のビデオ再生ウィンドウおよびコントロールがある。右側中ほどには、下部の時間バーに表示させる類似性スコアを選択するメニューコントロールがある。類似性スコアは、ビデオスライダバーと時間同期的に表示される。暗色領域は類似性の高い区間であり、濃くなるほど類似である。図は、表示されたフレームにあるように、暗い背景を背に中央にいる話者の中間クローズショットの類似性を示している。類似ショットの位置および程度は時間ラインの黒色バーで直接明らかとなる。

【0093】右側中ほどのスレッシュホールドスライダは、類似性スコアから索引点をどのように導き出すかを制御する。索引点は、時間バーの暗色（類似）領域の上部領域のやや明るいバーとして示されている。（この場合、これは主にB/W再現のためであり、索引点は類似性がスレッシュホールドを超えた時点で決定される。）時間バーの下に「|<<」および「>>|」のラベルが付けられたボタンは、再生点を次の索引点または前の索引点に自動的に進める。大きな類似性変動（多数の索引点）の領域では、ユーザは、スレッシュホールドを大きくすることによって最も重要な指標を選択できる。類似性が少ない領域では、ユーザは、スレッシュホールドを引き下げても索引点を見つけることができるが、信頼性が下がる。

【0094】図32は、本発明による、トレーニングビデオセグメントを対話的に定義し類似性測度を表示するための時間バーおよびユーザスレッシュホールドマウス入力を受け取るためのスレッシュホールドスライダバーを備えるブラウザを示している。時間バー3201は、類似であるとみられるビデオのセグメントを縦の黒色バーとして示す。スレッシュホールドスライダバー3202は、類似性の検出に必要な確率スレッシュホールドを指定するためのユーザのマウス入力を受け取る。時間バー3201は、例えばトレーニングセグメント指定についてクリック・ドラッグ操作によってユーザトレーニングマウス入力を受け取るように動作可能である。

【0095】図33は、ビデオの領域内のフレームを表示するためのスクロール可能ウィンドウ3301をさらに追加した図32のブラウザを示す。詳細には、メインブラウザウィンドウに表示され、時間バースライダ3303の位置によって指示されるフレーム3302およびその前後のフレームが、スクロール可能ウィンドウ3301に表示される。

【0096】このウェブ（Web）ベースのインタフェースは、極めて良好な概観を提供し、ビデオ全体の各種クラスをラベルづけるためのすぐれた選択となる一方で、ビデオ再生中の迅速な類似性探索のために特殊に仕上げられている。従って、水平スクロール可能ウィンドウ（図33の下部参照）に周期的にサンプリングされた類似の静止画像を示す追加表示が、本発明に従って選択的に含まれる。再生中、ウィンドウは、再生ウィンドウと同期して留まるように自動的にスクロールする。時間的脈絡は、再生ウィンドウに示されたフレームに最も近い静止画像をスクロール可能ウィンドウの中央に置くことによって示される。ビデオが停止されると、静止画像は誘導案内用を使用される。関心のある領域にスクロールさせ、その静止画像上でダブルクリックすると、ビデオが対応する時間のビデオに位置づけられる。

【0097】類似性探索の区間は静止画像上でマウスをドラッグすることによって選択される。選択された領域は、スクロール可能ウィンドウおよび時間バーの下部の両方に明緑色バーにより指示される。ビデオの小さな部分だけがスクロール可能ウィンドウの時間範囲内に表示されるので、示される選択領域はもっと大きなものである。図33で、スクロール可能ウィンドウに表示された選択領域は、スライダの爪のすぐ下のごく小さな領域に対応する。さらに、あらゆる時間依存媒体の場合と同様、ビデオに伴う問題は、何が選択されたのかが再生してみなければ必ずしも明白にならないということである。

【0098】類似性索引を作成するためには、最初に例題ビデオを選択しなければならない。1つのインタフェース方法は、ビデオの領域を選択するために図32および図33の時間バーで単純にクリック・ドラッグするこ

とである。あらゆる時間依存媒体の場合と同様、ビデオに伴う問題は、何が選択されたのかが再生してみなければ必ずしも明白にならないということである。前述の類似性測度の場合、最良の結果は、ソースビデオが、例えば同一のショットに由来するといったように、合理的に類似である場合に得られる。クリック・ドラッグ選択は、テキストの場合には効果的であるが、時としてユーザがほとんど気づかずに不要なビデオが選択される結果をもたらす。また、非接触選択も代替的に有効である。

【0099】図34は、1個以上のトレーニングビデオセグメントの終点を対話的に選択し、周期的フレームの類似性測度を表示するためにビデオの周期的フレームを表示するウェブベースのインタフェースを示す。ビデオ全体は最初に、図34に示されたように表示される周期的フレームに分割される。各周期的フレームは、ユーザがその周期的フレームを選択し、それをフレームセグメントに包含させるようにするチェックボックスを備える。隣接する周期的フレームがチェックされると、その2つのチェックされた周期的フレーム間の後続のビデオの全部の非表示フレームは、トレーニングセグメントの一部となる。例えば、周期的フレーム3401と周期的フレーム3402との間のビデオの全部のフレームはトレーニングセグメントに含まれる。ビデオの類似性探索が行われると、周期的フレームに対応する類似性情報は、周期的フレームの周囲の矩形ボックスの陰影として選択的に表示される。

【0100】図34は、選択された領域の視覚化と同時に非接触選択のサポートを可能にするビデオ領域選択用のウェブベースのアプリケーションを示している。このアプリケーションでは、ビデオは、通常の区間で切り取られた一連のキーフレームとして表される。図34は、選択された領域の視覚化と同時に非接触選択のサポートを可能にするビデオ領域選択用のウェブベースのアプリケーションを示している。このアプリケーションでは、ビデオは、通常の区間として切り取られた一連のキーフレームとして表され、それらのビデオにおける時間（秒単位）とともに示される。ビデオ録画プレゼンテーションの場合には5秒間隔が適切であるが、他の用途ではそれより速いかまたは遅いレートも選択的に好適である。ユーザは、各フレームの下をチェックボックスをクリックすることによって複数のキーフレームを選択する。隣接して選択されたキーフレーム間のビデオの全フレームについてモデルがトレーニングされる。このインタフェースは、終点を精確に位置決め可能とし、選択されたビデオ内容を明示的に表示するという理由で、クリック・ドラッグよりもある点ですぐれている。また図34は、非接触選択が複数の区間を次々と選択することにより可能であることも示している。このインタフェースは、簡潔な表示により、ユーザが一目で関心のある領域を見つけられるようにする。通常サイズのウェブブラウザで

は、10分のビデオに対応する120個の画像がウィンドウに示され、残りのビデオもスクロールによって容易にアクセス可能である。インタフェースは、様々なクラスの画像への様々なラベルの割り当てもサポートする。以前に割り当てられたラベルは表示ではカラーコード化される。選択されたビデオの類似性は、ほぼ即時的に計算され、図32および図33のブラウザに表示されるか、または、スレッシュホールドで切られ、図34のように各フレームの周囲に異なる色でウェブインタフェースに表示される。

【0101】図35は、本発明に従って離散コサイン変換およびアダマール変換係数によって計算されたビデオの類似性マトリックスを示す。距離計量の利用を示すために、全部のフレーム間の類似性を計算し、結果のマトリックスを画像として表示することにより、ビデオの自己類似性を視覚化することができる。図35は、スタッフ会議のビデオの距離マトリックスを示す。位置(i, j)の各画素は、類似フレームであればあるほど色濃くなるように、フレームiとフレームjとの間の距離に比例して着色されている。各軸の単位は秒単位での時間であり、各点は、最高分散を有する100個の離散コサイン変換およびアダマール変換係数間のユークリッド距離に比例して着色されている。アダマール変換ドメインに関して従来しばしばなされた批判は、知覚的相違と良好に相関しないということである。アダマール変換は一般にクラスタ化およびモデル化について同様に良好に作用するが、距離がアダマール変換および離散コサイン変換の両方の表現に関して極めて類似であることを指摘しておくことは興味深い。i=jにおける黒色直交線は、フレームがそれら自身と同一であることを指示する。いくつかの特徴が目につき、後続部分と類似でないビデオの始まりの導入期間が存在し、それは約500秒続くことが容易にわかる。

【0102】右下隅の4個の濃色の正方形は、スライドプレゼンテーションの2つのロングショットに由来する。個々のスライドの変化はそこに見ることができるが、それらは聴衆または話者のカットよりも小さい大きさのものである。これらのスライドは、約550秒に開始する別のスライドプレゼンテーションとも極めて類似であり、同じく自己類似である聴衆のショットとインターカットし、「チェッカーボード」パターンを生じる。またスライドは、1600秒および1900秒のコンピュータデスクトップのショットともある程度類似であり、それらの領域を濃色に見せているが、他のスライド領域ほど濃くはない。これらのマトリックスは全体的に直観的ではなく、いずれかの特定の時間に得られる「スライス」は、ビデオの残部に対するその時間におけるそのフレームの類似性を示している。図32および図33の時間バーとして提示されると、これは、単一のフレームが類似のビデオ領域を見つけるためにどのように

使用されるかを示すが、ガウスモデルは、分散をモデル化できるためによりロバストである傾向がある。

【0103】本発明はまた、カラー情報にもとづき1個以上の付加的なシグネーチャを計算することによって、カラー検索を行うための改良を含む。これは、特徴ベクトルによって表現される現行の輝度(Y)シグネーチャに付加するために画像の色成分(YUV色空間におけるUV成分)に関する付加的な特徴シグネーチャを計算することによって実現される。色成分は少ない空間解像度を要するので、それらは少ないシグネーチャで表現される。本質的に、フレームの色成分の変換からの変換係数位置が選択され、特徴ベクトルに追加され、それにより、特徴ベクトルは同一カラーフレームから得られた輝度フレームおよび色フレームの両方の変換からの係数を含む。

【0104】別の代替法によれば、YUBまたはRGBの各カラー成分は個別の画像フレームとして扱われる。従って、各フレームに対して3つの変換が適用され、シグネーチャ(特徴ベクトル)は各個別画像について計算されて比較される。これは、類似性計量における全カラーによる重みづけを可能にする。カラー情報の包含のための本発明に従ったさらに別の代替法は、この検索技法と別の、例えばカラーヒストグラムにもとづく技法との組合せである。初期の類似性工程において、画像は輝度特徴ベクトルによって類似性がわかる。その画像を領域に分解し、各領域についてカラーヒストグラムを計算することによって、画像における空間情報の一部が保存される。最終類似性工程では、初期類似性工程から得られた最上位画像が、カラーヒストグラム類似性評価法または他の類似性評価法によって類似性について再度スコアが付けられる。

【0105】カラーは、多くの種類のビデオ画像にとって、例えばコンピュータプレゼンテーションがスライドの背景色だけで識別できる場合が多いスタッフ会議のビデオにおいて、有効な糸口である。また、動きまたは時間シーケンスのモデル化も多くの用途で極めて有用であり、より強力な統計モデルがそれを可能にする。

【0106】ガウスモデルは多くの用途にとって有効であるが、区間内の全部の変化が平均化されるという短所を有する。時間的シーケンスまたは継続時間を捕捉することが重要である場合、隠れマルコフモデルが代替的に使用される。隠れマルコフモデルの出力分布は、まさしく前述の通り、特徴ベクトル上の単数または複数のガウスモデルとしてモデル化される。隠れマルコフモデルの利点は、各状態が暗示的または明示的な継続時間モデルを有することである。これは、(過度に長いまたは短い)ありそうにもない継続時間のショットにペナルティーを科す因子を尤度計算に加える。これは、継続時間モデルが同一状態と最も隣接するフレームを拘束し、従って擬似的なショット境界を低減するので、単純な最大尤

度フレーム分類よりも有効である。

【0107】隠れマルコフモデルでの公式化は、複数の状態および(音声認識における言語モデルに類似の)遷移グラフを用いてビデオの傾向またはシーケンスを捕捉するために選択的に強力に拡張される。従って、隠れマルコフモデルは、例えば、ニュース放送の開始を特徴づける放送局のロゴからニュースキャスターのショットへの遷移をモデル化するために選択的に使用される。隠れマルコフモデルに内在するシーケンス拘束のために、これは、放送の終了時に多く生じるニュースキャスターのショットから放送局のロゴへの遷移には一致しないが、単純ガウスモデルは両者の場合について高いスコアを生じる。

【0108】また、元の特徴ベクトルのフレーム間差異として計算される差分表現も有用である。パーセプualの関係によって、各ベクトルのノルムは、画素の差のノルムに(ほぼ)比例する。従って、カットまたはカメラの移動によって生じた大きなフレーム間差異は、差分ベクトルのノルムを計算することによって容易に検出される。あるいはまた、それらは、動きを捕捉する追加の特徴を形成するために元の特徴ベクトルと連結される。

【0109】本発明に従った類似性探索の方法は、類似のビデオ領域を見つける迅速かつ強力な手段を記述する。ユーザが例題ビデオを用いてクエリーを指定できるようにすることは、テキストベースまたはスケッチベースのインタフェースを凌ぐ進歩である。この技法は、大きなビデオコレクションに、さらにカラーまたは時間的類似性の測度に容易に拡張される。

【0110】週毎のスタッフ会議が、複数のビデオカメラおよびマイクロフォンが装備された会議室で開かれることもある。会議は、経営陣およびスタッフによる全体発表に始まり、その後個々の職員によるプレゼンテーションに進む。プレゼンテーションは通常1人によって行われ、オーバヘッドプロジェクタまたはコンピュータによるスライドといったグラフィックスを含み、一般に会議では1つ以上のプレゼンテーションが行われる。カメラ担当者は、部屋のカメラを切換え、ビデオ録画のショットを提示する。ビデオはMPEG符号化され、社内イントラネットによってスタッフに利用可能となる。

【0111】図36は、本発明に従ったオーディオ・ビジュアル記録物をセグメント化する方法に対応するデータの流れを示す。ソースビデオ3601は工程3602でスライド領域を見つけるために分析される。ソースビデオ3601のオーディオチャネルは、スライド区間に対応するソースビデオ3601の領域について工程3603で抽出される。工程3603で抽出されたオーディオ区間は、話者ごとに工程3604でクラスタ化される。すなわち、オーディオ区間は、相互に比較照合され、それらのソースに従って分類される。得られたオーディオ区間のクラスタは、各々が単一話者に由来するも

のとみなされる。同一話者クラスタのオーディオ区間は工程3605で併合される。工程3606で、ソース特定話者モデルが各併合オーディオ区間についてトレーニングされる。工程3607で、ソースビデオ3601のオーディオチャンネルは、話者認識によって話者ごとにセグメント化される。オーディオチャンネルによるセグメント化の結果は、以後のブラウジングおよびソース特定検索操作のためにソースビデオ3601およびソースオーディオ3608において索引づけられる。

【0112】図37は、2人の話者による2つのプレゼンテーションを有する記録された会議のスライドであるオーディオ・ビジュアル記録物のフレームの確率の対数を示す。話者Aのプレゼンテーションの範囲を示すラベル3701は、ビデオを見ている人間のユーザにより得られた話者Aのプレゼンテーションの実際に観測された継続時間である。同様に、話者Bの指標3702は話者Bのプレゼンテーションの全範囲を示す。

【0113】各フレームのコンパクトな特徴ベクトル（簡約化された係数）が上述の通り計算される。対角共分散ガウスモデルは、いくつかの無関係な会議ビデオからのスライド画像でトレーニングされている。このモデル

フレームごとのプレゼンテーション分類誤り

使用された特徴	検出漏れ	検出誤り
スライド	0.745	0.058
スライド+話者セグメント化	0.042	0.013

【0115】図38は、図36に示した工程3604および3605に示したような本発明に従ったオーディオ区間に適用されるクラスタ化方法におけるデータの流れを示す。オーディオ区間3801～3804は、図36に示したソースオーディオ3608から抽出された、図37で1、2、3および4のラベルが付けられた4個のオーディオ区間を表している。オーディオ区間3801～3804はオーディオベクトル3805～3808にパラメータ化される。クラスタ化法3809がオーディオベクトル3805～3808に適用され、相互に小さいユークリッド距離を有するオーディオベクトルに集塊させる。クラスタ化法3809の結果は、それぞれ話者AおよびBに対応するオーディオ区間3810およびオーディオ区間3811と併合される。

【0116】ある話者の口から数センチメートル以上離れたファーストフィールドマイクロフォンによって話者識別を行うことは特に困難である。記録された会議でのオーディオは演壇マイクロフォンまたは他のクローズトーンマイクロフォンではなく複数の天井マイクロフォンから得られるので、話者識別は特に困難になる。実際にあらゆる話者識別技法は、特定の話者を特徴づけるためにメル周波数ケプストラル係数 (mel-frequency cepstral coefficient) といった何らかの種類のオーディオスペクトル測度を使用する。あらゆる現実的環境におけるファースト

\*ルは、各ビデオフレームに関する尤度を生成するために使用され、それはそのフレームがスライドであるという対数尤度を測定する。1個の標準偏差をスレッシュホールドとした場合、そのビデオにおいてスライドが表示された時点の確実な評価値を生じる。下記の表3に示すように、スライドは94%の確度でプレゼンテーションと関係づけられた。20秒以上の長さのスライド区間がシステムの候補スピーチ区間として使用される。図37は、スタッフ会議のスライドの対数尤度のプロットを示している。20秒以上の長さの上記のスレッシュホールド（点線）である判定基準を満たす4個の区間が存在し、それらは1、2、3および4のラベルが付けられている。この特定の会議において、それぞれAおよびBのラベルが付けられた2人の話者により行われた2つのプレゼンテーションが存在した。各プレゼンテーションの範囲は図37の上部に示されており、それはセグメント化実験に関する実地検証情報として機能する。話者Bのプレゼンテーションは、スライドが表示された期間の2倍以上続けられたことに留意されたい。

【0114】

【表3】

マイクロフォンは、直接的に、また、壁、床、机といった環境配置によって反射された音声を拾ってしまう。こうしたマルチパス反射は、音声の周波数スペクトルを著しく変更するくし形フィルタ効果をもたらす。この問題は、（遠隔会議システムにおいて普通に行われているように）複数のマイクロフォンからの信号を混合することによってさらに悪化する。部屋の共鳴による付加的な効果も各マイクロフォンの周波数応答に影響する。共鳴およびくし形フィルタ効果はともに、室内の話者の位置により著しくかつ予測不可能に変化する。これは、トレーニングスピーチのサンプルを使用して話者モデルをトレーニングする現在の話者識別法を、ファーストフィールドマイクロフォン環境にとって特に不適にさせる。音響環境によるスペクトル変化はしばしば、話者間のスペクトル差異とほとんど同じ程度の大きさである。

【0117】予測できない室内音響によるトレーニングデータと試験データとの間の不可避免的な不一致を回避するために、本システムは本質的に、単一話者によって発せられたと思えるセグメントを抽出することによって試験データからトレーニングデータを取得する。現在の実施の形態において、これは、単一話者のスピーチがスライドといったプレゼンテーション視覚物の表示と関連していると仮定することによって行われる。（仮定されたスタッフ会議の領域分野では、この仮定は、完全にはないが通常は、所与のスライド区間において質問、笑声

または他の感嘆が頻繁に存在するので、正確である。)

【0118】単純な顔面またはニュースキャスター検出といった他のビデオ分析は同様に使用される。本発明に従った代替法として、顔面認識は、ビデオ区間を特定の話者と関係づけるために使用されるオーディオクラスタ化を強化または代替できる。

【0119】次の工程は、何人の話者がスライドプレゼンテーションを行ったかを判定するために候補区間をクラスタ化することである。これは、任意の数のクラスタ化技法のいずれかによって行えるが、現在の実施の形態の場合、オーディオ類似性の極めて単純な測度が使用される。各オーディオ区間はメル周波数ケプストラル係数にパラメータ化され、各区間の係数の平均が比較照合される。ユークリッド距離測度および、最大距離の1/2をスレッシュホールドとする集塊クラスタ化法によって、各話者候補に関する個別のクラスタが得られる。クラスタ化スレッシュホールドは、いずれかの既存のクラスタに十分に類似でない区間を排除する。例えば、あるスライドに関するクエリがなされる場合、得られる区間はほとんど、多数の異なる話者からのスピーチを含む。より精緻な距離およびクラスタ化法、例えば、ノンパラメトリック類似性測度、尤度比距離および/または可変スレッシュホールドクラスタ化といった方法が選択的に使用される。隣接セグメントのクラスタ化を助成するために距離測度にバイアスをかけるといった付加的な拘束または、話者の数に関する事前の知識を使用することにより、選択的にクラスタ化を改善させることもできる。前述の通り、自動顔面認識は音響クラスタ化を代替的に強化または代替できる。

【0120】図39は、本発明に従った一連の話者単位より構成される話者遷移モデルを示す。フィラーモデル3901、3903および3905は、例えばビデオの非単一話者セグメントでトレーニングされるオーディオモデルを表す。話者モデル3904は、図38に示した併合オーディオ区間3810でトレーニングされる話者モデルを表す。話者モデル3905は、図38に示した併合オーディオ区間3811でトレーニングされるモデルを表す。話者単位3806および3907は、セグメント化における話者シーケンスの知識によってソースオーディオ3608をセグメント化するために図36に示す工程3607で使用される隠れマルコフモデルを形成するために連結される。

【0121】クラスタ化の結果から、プレゼンテーションを行う話者の数および彼らが話す順番が決定される。これは隠れマルコフモデルを用いてビデオをセグメント化できるようにする。さらに、クラスタ化されたオーディオセグメントは各話者モデルをトレーニングするために使用される。クラスタ化の結果から、ビデオの時間範囲をモデル化するために隠れマルコフモデルが自動的に構築される。図39はモデルの構造を示している。「フ

ィラー」モデルは、発表者の話以外とみなされるオーディオを表す。この実施の形態では、フィラーモデルは、ソースビデオの最初の2分間からのオーディオと同様、他の会議ビデオからセグメント化された沈黙、笑声、称賛および聴衆の雑音でトレーニングされ、それはプレゼンテーションの話者による話を含まないといみなされる。フィラーモデルは、多重事例化されているが、好ましくは各事例で同一である。話者特定モデルはプレゼンテーションの話者からの話を表す。各話者特定モデルは、それに関係する結合されたスライド区間のクラスタからのオーディオでトレーニングされる。話者モデルおよび選択的なフィラーモデルを連結することにより「話者単位」が得られる。それらは、話者ごとに1個ずつ連結され、最終モデルを生じる。これにより正しい話者シーケンスが得られる。セグメント化は、完全モデルによりソースオーディオの最大尤度アライメントを見つけるためにバイタービアルゴリズムによって実行される。これは、スライドが表示される区間と実質的には異なる可能性があるので、各発表者の話の範囲を決定可能にする。特に、話者が話している間に話者のショット、聴衆のショットおよびプレゼンテーションスライドの間で交替が起こることはビデオにとって普通である。この実施の形態では、フィラーモデルおよび話者モデルともに単一の状態を有しており、単一混合の全共分散ガウス出力分布を有する。モデルが単一状態および単一混合を有するので、それらは1パスで迅速にトレーニングされる。複数状態または複数混合モデルは、より高価なトレーニングによって性能を改善できよう。自己遷移はいかなるペナルティも伴わずに可能であり、明示的な時間継続をいっさい持たないエルゴード的モデルを生じる。これにより、モデルは、いかなる確率ペナルティも伴わずに所与の時間長を表現することができる。

【0122】図40は、本発明によるオーディオ・ビジュアル記録物をセグメント化する方法のセグメント化の結果を例示している。このように、話者Aの指標4001は、話者Aのプレゼンテーションの実際の継続時間4003にほぼ重なり合っている話者Aのセグメント化を表す。話者Bのセグメント化指標4002は、セグメント化が実際の話者Bの継続時間4004にほぼ重なり合う結果となったことを表す。このようにして、話者Aの指標4001および話者Bの指標4002は、本発明によるセグメント化によって作成される索引より導出される。

【0123】図40は、会議のソースビデオに関する自動セグメント化の結果を示す。不利な音響環境(利得制御を伴う6個のファーフールドマイクロフォン)にもかかわらず、2人の話者は識別され、彼らのプレゼンテーションの範囲は、数十秒以内まで合理的に良好にセグメント化された。これはビデオのセグメント化およびブラウズにとって明らかに妥当である。最大の不一致は話

者Aのプレゼンテーションの終わりにあり、それは事実上話者Bのプレゼンテーションの開始まで続くようにセグメント化された。これはたぶん、2人の話者が、映写装置の詳細を話し合っていたのでその区間に話をしていたためであろう。

【0124】単一の会議を選択するために使用される同じ技法は、同じ話者の組を含む複数の会議に対しても選択的に適用される。個々の会議からのプレゼンテーションは会議の資料について選択的にクラスタ化される。これは発表者の目録を作成可能にする。それが潜在的に異なる音響環境（部屋の位置）における同一話者の話の十分な実例を含んでいれば、より強固な、位置に依存しない話者モデルが選択的にトレーニングされる。さらに、会議進行表において話者が識別されていれば、話者モデルは以後の識別および検索のために氏名と関係づけられる。

【0125】スライドプレゼンテーションを含む6本のビデオ録画された会議が試験資料として使用された。オーディオフィラーモデルおよびスライド画像のトレーニングデータは別の組のビデオから得た。6本のビデオの合計長さは280分21秒であり、約45分の平均長であった。各ビデオは1～5本のプレゼンテーションを含み、合計16本であったが、3本のプレゼンテーションはビデオおよびスライドを含んでおり、ほとんどが聴衆の質問または注釈を有していた。プレゼンテーションは一般にスライド区間の継続時間より長いので、スライドの存在はプレゼンテーションの良好な指標であり、スライドだけからプレゼンテーションを見つけることはプレゼンテーションの75%を見逃す結果となった。表3の第2行は、話者のセグメント化がこれをどれほど改善さ

フレームごとのプレゼンテーション分類誤り

終点検出	リコール	精度
クラスタ化前	0.81	0.23
クラスタ化後	0.81	0.57

【0128】本発明によるこれらの方法は、会議ビデオの他に、個々の話者が識別可能なビデオ特徴に関係づけられるあらゆる分野に適用可能である。一例は、ニュースキャスターのショットが画像構成および背景により識別できる場合が多い、ニュース放送である。話者識別の使用により、ロケーションまたは他の介在ビデオが存在する場合でも、ニュースキャスターによるニュース記事のセグメント化が可能である。

【0129】図41は、本発明に従ったセグメント間音響距離マトリックスを示す。対角線上成分4101～4105は、各セグメントがそれ自体に類似であることを示す黒色である。灰色領域4106および4107は、ソースオーディオの始まりおよび終わりにおけるオーディオ区間の部分的類似性を表す。白色領域はオーディオセグメントの非類似を表す。

【0130】多くの場合、例えば図40でラベル2、3

\*せるかを示す。プレゼンテーションの約5%だけがプレゼンテーション以外のものであると誤って識別された。

【0126】16本のプレゼンテーションにもとづき、（ビデオおよび変則的なオーディオによる付加的な終点とともに）合計32個の検出すべき終点が存在した。実際の話者の話の開始または終了の15秒以内に生じていれば、終点は正確であるとみなした。表4は終点の位置の確度を示す。クラスタ化以前に、57のスライド区間による114個の終点が存在した。検出すべき32個の関連する終点の实地検証情報が与えられ、26個の終点が正確に突き止められて、これは0.23の精度による0.81のリコールをもたらし、ほとんどの終点は見つかったが、それが正しい終点である可能性が1/4未満であることを意味する。57個のアライメントされたセグメントをクラスタ化することにより23個のクラスタを得たが、これは不正確な終点の数を減らすことにより精度を劇的に改善させた。検出された終点のうち少なくとも2個はプレゼンテーションに対するビデオ区間によっており、精度は不当に悲観的であることに留意されたい。非理想的オーディオ環境もクラスタ化問題を生じた。マイクロフォンはHVACベント付近の音響天井タイルに設置されている。いくつかのプレゼンテーションは換気雑音の有無により誤ってクラスタ化された。これは音響信号に大きな影響を与え、同じ話者も換気システムの状態によって別様にクラスタ化され、一部のクラスタ境界はまさに換気スイッチのオンオフにより生じている。

【0127】

【表4】

および4が付けられたような、同一話者に対応する複数の隣接区間が存在する。クラスタ化は、尤度比距離などの多くの技法によって代替的に実行される。ここで使用するクラスタ化法は、ノンパラメトリック距離測度にもとづく。オーディオセグメントにパラメータ化されたメル周波数ケプストラル成分は、クラス境界を見つけるために最大相互情報量評価基準を用いて監視ベクトル量子化数をトレーニングするために使用される。トレーニングされると、セグメントはベクトル量子化され、二項分布のヒストグラムが作成される。このヒストグラムは、オーディオフィイルのシグネチャとして機能し、ベクトルとして処理される場合には2つのヒストグラム間のコサインはオーディオ類似性の良好な測度として機能する。図41はこの測度を用いて計算された距離マトリックスを示す。これは、単一の会議ビデオからの12個のスライド領域の間のオーディオ類似性を示している。各

成分  $i, j$  は、より近い距離、すなわちより類似性であるものが濃色になるように、セグメント  $i$  および  $j$  の間の距離を図示するように着色されている。図41から、各々が特定の話者による話に対応する、いくつかの音響的に類似の群が存在することは明白である。例外は、中央の話者のプレゼンテーションにおいて示されたビデオからのタイトルに対応する、セグメント7によるものである。このような距離マトリックスは、単一話者に対応する類似区間を見つけるためにクラスタ化される。いずれかの種類の階層的クラスタ化が選択的に使用される

が、ここで採った単純な方式は、各自の距離のいずれもスレッシュホールドを超えない限り、全部の隣接セグメントを同一クラスタの一部であるとみなすことによって、クラスタメンバーの時間隣接性を強制することであった。図41のセグメントの場合、これは以下のように5個のクラスタとなった。

(1, 2, 3, 4, 5) --- (6) --- (7) --- (8) --- (9, 10, 11, 12)

【0131】 実地検証情報は3つのプレゼンテーションが存在するというものであったので、このクラスタ化方法は、第2のプレゼンテーションを、オーディオ距離にもとづき3個に誤ってセグメント化した。重要な目的はビデオブラウジングのための索引を見つけることなので、それは絶望的な誤りではない。プレゼンテーションが開始した時点と同様、ビデオが表示された時点を見つけることも望ましい。より精緻なクラスタ化方法は、図41のセグメント7といったオーディオアウトライアーまたは、質問や称賛といった他の変則的オーディオを無視するために使用される。

【0132】 セグメント化プロセスにおける第1工程は、ビデオにおけるスライドを突き止めることである。これは、プレゼンテーショングラフィックスがそのビデオにおいて表示される時点の正確な推定値をもたらす、上述の本発明による技法によって行われる。元のMPEG-1ビデオは、時間に関して2フレーム/秒に、空間に関して64×64画素表現の下位画像に間引かれる。各簡約化されたフレームはその後、離散コサイン変換またはアダマール変換によって変換される。変換は、画像圧縮の場合に普通である小さな下位ブロックに対してではなく、フレーム画像全体に適用される。変換されたデータはその後、その100個の主成分に射影により簡約化される。

【0133】 図42は、本発明に従って、スライドビデオ画像と類似である所定の時間間隔よりも長い1個以上のビデオフレーム区間を識別する方法を示している。工程4201で、ビデオは時間および空間に関して間引かれる。工程4202で、フレームは離散コサイン変換またはアダマール変換によって変換される。工程4203では、工程4202で計算された変換マトリックスから特徴ベクトルが抽出される。工程4204で、スライド

の確率がスライド画像クラスのガウスモデルを用いて計算される。工程4205では、工程4204において計算された尤度が、そのフレームがスライド画像クラスと類似であるか否かを判定するためにスレッシュホールドと比較される。それがスライドであると判定されると、工程4206は、以前のNフレームもスライドであったかどうかを検査する。Nは、工程4207でスライド区間が見つかる前に、検出されるスライドの所定の時間間隔が超えられなければならないように選択される。例えば、20秒のスライドスレッシュホールドで、2フレーム/秒に間引く場合、Nは40であるように選択される。従って、単一フレームがスライドであると判定されたが、そのスライドフレーム以前のフレームおよびスライドフレーム以降のフレームがスライドでなければ、スライド区間はラベルづけされない。工程4205がそのフレームは非スライドであると判定した場合または現在のフレームはスライドであるが以前のNフレームはスライドではないと判定した場合、工程4208は、ビデオの終わりに到達したかどうかを検査する。さらにフレームがある場合、方法は再び工程4202からその次のフレームに対して開始する。ビデオの終わりに到達していれば、方法は図43に進む。

【0134】 図43は、本発明に従ったスライド区間から抽出されたオーディオ区間によるソース特定話者モデルをトレーニングする方法を示している。工程4301で、スライド区間に対応するオーディオ区間が抽出される。この抽出は、そのスライド区間が抽出されたソースビデオ3601に対応する図36に示したソースオーディオ3608により行われる。工程4302で、最初のオーディオ区間がメル周波数ケプストラル係数にパラメータ化される。オーディオ区間に対応する多様なメル周波数ケプストラル係数ベクトルは、そのオーディオ区間に対応するオーディオ係数平均ベクトルを生成するために工程4303で平均化される。さらにオーディオ区間があれば、工程4304は、次のオーディオ区間の処理のために方法を工程4302に戻す。全部のオーディオ区間がパラメータ化され、オーディオ係数平均ベクトルが各オーディオ区間について計算されると、オーディオ区間は工程4305でクラスタ化される。工程4305は同一話者判定基準によってオーディオ区間をクラスタ化する。すなわち、ユークリッド距離に関して相互に十分に近いオーディオ係数平均ベクトルを有するオーディオ区間は、同一話者によるものであると判断される。工程4306で、同一クラスタのオーディオ区間が併合される。工程4307で、第1の話者モデルが第1の併合オーディオ区間でトレーニングされる。検査4308は、併合オーディオ区間のクラスタがさらに存在するかどうか判断される。肯定であれば、工程4307は、一意的に決まる話者モデルをトレーニングするために全部の併合オーディオ区間が使用されるまで次々に処理す

る。

【0135】図44は、本発明に従った話者遷移モデルを用いてオーディオ・ビジュアル記録物をセグメント化する方法を示す。工程4401で、オーディオの隠れマルコフモデルが構築される。図39は、工程4401によって構築されるようなオーディオ隠れマルコフモデルを示している。ビデオおよびオーディオは、工程4402でそのオーディオ隠れマルコフモデルによってセグメント化される。工程4403で、ビデオおよびオーディオは、工程4402で決定されたセグメント化情報により索引づけられる。このように、図44に示す方法は、図36に示した工程3607を実施するために適する。

【0136】会議の進行表が得られる場合、プレゼンテーションは、進行表からの情報を用いて選択的に自動的にラベルづけまたは索引づけされる。これにより、プレゼンテーションは発表者および演題によって容易に見つけることができる。このようにして、会議ビデオは、内容によって自動的に索引づけ、ブラウジングおよび検索される。

【0137】本発明をいくつかの態様および実施の形態に関して説明したが、これらの態様および実施の形態は、限定としてではなく、例示として提起されている。本発明の精神および範囲を逸脱することなく各種の追加および変更が行い得ることを理解しなければならない。例えば、数倍の改善といった精緻な音響モデルは、継続時間モデルを各話者に対して強制することによって代替的に得られる。別の例として、オーディオ特徴と同様にビデオ特徴にもとづくセグメントのクラスタ化は、発表者のスライドが、発表者自身の画像だけでなく、類似性の構成およびカラー図式を有するはずであるという仮定にもとづき、本発明に包含される。それにより、オーディオおよびビデオの両方の変則的領域の識別をプレゼンテーション中に表示されるビデオによって可能にする。また別の例として、対話的に定義された探索セグメントを指定するユーザ入力を受け取るための他のウェブベースのインタフェースが使用できる。さらに別の例として、ガウス分布以外の確率分布を用いた分類が適切な状況において使用することができる。従って、こうした追加および変更はすべて、特許請求の範囲に記載された本発明の精神および範囲に通じるものであると見なされるべきである。

#### 【図面の簡単な説明】

【図1】本発明の方法を実行するために適した汎用コンピュータアーキテクチャを示す。

【図2】本発明によるビデオの分類を実行する方法におけるデータの流れを示す。

【図3】本発明による、トレーニングフレーム、トレーニングフレームから得られた平均特徴ベクトルの逆離散コサイン変換およびトレーニングフレームから得られた平均特徴ベクトルの逆アダマール変換を示す。

【図4】異なる平均および分散を有する一次元ガウス分布を示すグラフである。

【図5】本発明によるビデオ分類のための特徴集合を選択する方法を示すフローチャートである。

【図6】ビデオフレームの離散コサイン変換により得られる変換マトリックスを示す。

【図7】本発明に従って2個以上の変換マトリックスから計算された分散マトリックスを示す。

【図8】本発明に従って切り捨てによって決定された特徴集合を示す。

【図9】本発明による図8に示した特徴集合を有するトレーニングフレームの2個以上の特徴ベクトルから計算された平均特徴ベクトルを示す。

【図10】本発明による図8に示した特徴集合を有するトレーニングフレームの2個以上の特徴ベクトルから計算された対角共分散マトリックスを示す。

【図11】本発明の方法に従って分類のために図8に示した特徴集合を有するフレームについて検索された特徴ベクトルを示す。

【図12】本発明に従って2個以上のビデオ画像クラスのいずれかにビデオのフレームを分類する方法を示すフローチャートである。

【図13】本発明に従って、主成分分析、最大分散を有する係数の選択または最大平均を有する係数の選択により決定された特徴集合を示す。

【図14】本発明による図13に示した特徴集合を有するトレーニングフレームの2個以上の特徴ベクトルから計算された平均特徴ベクトルを示す。

【図15】本発明による図13に示した特徴集合を有するトレーニングフレームの2個以上の特徴ベクトルから計算された対角共分散マトリックスを示す。

【図16】本発明の方法に従って分類のために図13に示した特徴集合を有するフレームについて検索された特徴ベクトルを示す。

【図17】本発明による類似性を決定する方法において、類似性を決定するためのスレッシュホールドとして使用されるスライド画像クラス統計モデルの標準偏差の倍数の関数として、スライドとして正確に識別されたスライドフレームの割合およびスライドとして誤って識別された非スライドフレームの割合を示すグラフである。

【図18】本発明に従って画像クラス統計モデルを用いてビデオフレームの類似性を決定する方法を示すフローチャートである。

【図19】本発明に従ってビデオの各種フレームを生成するビデオ画像クラス統計モデルの確率の対数の表示を示すグラフである。

【図20】本発明に従ってビデオの各種フレームを生成するビデオ画像クラス統計モデルの確率の対数を表示する方法を示すフローチャートである。

【図21】特徴集合の成分の数dの関数として正確に分

類されたフレームの割合、それらのフレームに適用された変換のタイプおよびd成分特徴集合の選択方法を示すグラフである。

【図22】本発明の方法に従ってスライドと類似と見られるビデオの領域を表示するブラウザを示す。

【図23】本発明によるビデオを分類する方法において使用される隠れマルコフモデルに対応するクラス遷移図を示す。

【図24】図23に示すクラス遷移図に対応する本発明に従ったクラス遷移確率マトリックスを示す。

【図25】図23に示すクラス遷移図に従った5連続初期ビデオフレームに対応する全部の可能なクラスシーケンスを示す。

【図26】本発明に従ってクラス遷移確率マトリックスおよび画像クラス統計モデルを用いたビデオをセグメント化する方法を示すフローチャートである。

【図27】本発明による類似性探索を実行する方法におけるデータの流れを示す。

【図28】本発明によるビデオに対応する特徴ベクトルデータベースを計算する方法を示すフローチャートである。

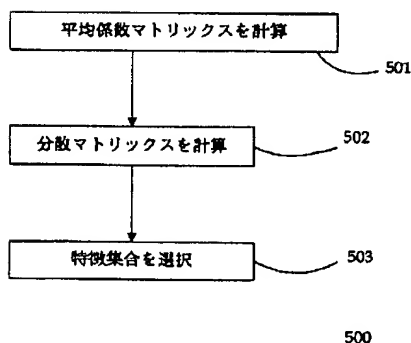
【図29】本発明による統計モデルを対話的にトレーニングする方法を示すフローチャートである。

【図30】本発明によるブラウザ内でビデオフレームを提示し類似性測定を表示する方法を示すフローチャートである。

【図31】本発明に従って、対話的に定義されたトレーニングビデオセグメント、そのトレーニングビデオセグメントのトレーニングフレームから得られた平均特徴ベクトルの逆離散コサイン変換および、トレーニングビデオセグメントのトレーニングフレームから得られた平均特徴ベクトルの逆アダマル変換を示す。

【図32】本発明による、トレーニングビデオセグメントを対話的に定義し類似性測定を表示する時間バーおよび、ユーザスレッシュホールドマウス入力を受け取るスレッシュホールドスライダバーを備えるブラウザを示す。

【図5】



500

【図13】

1	$C_{11}$
2	$C_{12}$
3	$C_{21}$
4	$C_{23}$
5	$C_{25}$
6	$C_{32}$

1300

【図14】

1	$\mu_{11}$
2	$\mu_{12}$
3	$\mu_{21}$
4	$\mu_{23}$
5	$\mu_{25}$
6	$\mu_{32}$

1400

【図15】

	1	2	3	4	5	6
1	$V_{11}$	0	0	0	0	0
2	0	$V_{12}$	0	0	0	0
3	0	0	$V_{21}$	0	0	0
4	0	0	0	$V_{23}$	0	0
5	0	0	0	0	$V_{25}$	0
6	0	0	0	0	0	$V_{32}$

1500

【図33】ビデオの領域内にフレームを表示するためのスクロール可能ウィンドウが追加された図32のブラウザを示す。

【図34】1個以上のトレーニングビデオセグメントの終点を対話的に選択し、周期的フレームの類似性測定を表示する、ビデオの周期的フレームを表示するウェブベースのインタフェースを示す。

【図35】本発明に従って離散コサイン変換係数およびアダマル変換係数を用いて計算されたビデオの類似性マトリックスを示す。

【図36】本発明によるオーディオ・ビジュアル記録物をセグメント化する方法に対応するデータの流れを示す。

【図37】2人の話者による2つのプレゼンテーションを含む記録された会議のスライドであるオーディオ・ビジュアル記録物のフレームの確率の対数を示すグラフである。

【図38】本発明によるオーディオ区間に適用されるクラスタ化方法におけるデータの流れを示す。

【図39】本発明による一連の話者単位を構成する話者遷移モデルを示す。

【図40】本発明によるオーディオ・ビジュアル記録物をセグメント化する方法のセグメント化結果を示すグラフである。

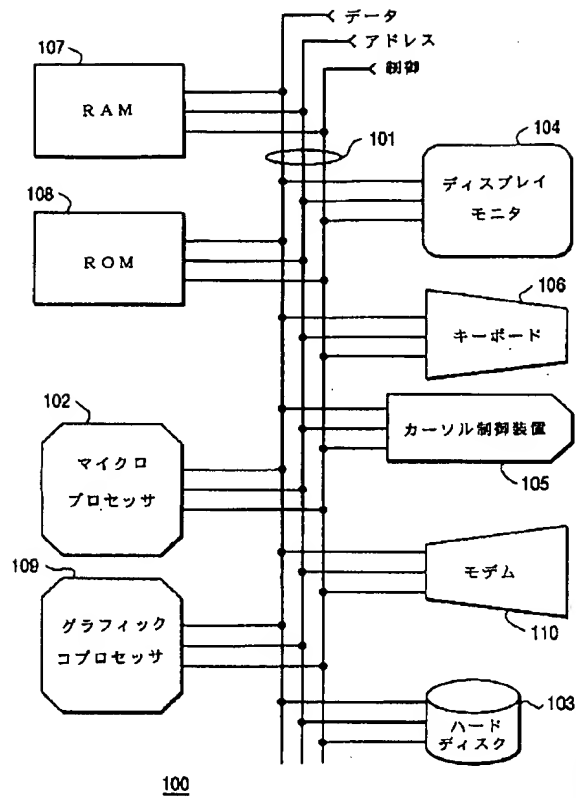
【図41】本発明によるセグメント間音響距離マトリックスを示す。

【図42】本発明による、スライド画像クラスとの類似性を有する所定の時間間隔より長い1個以上のビデオフレーム区間を識別する方法を示すフローチャートである。

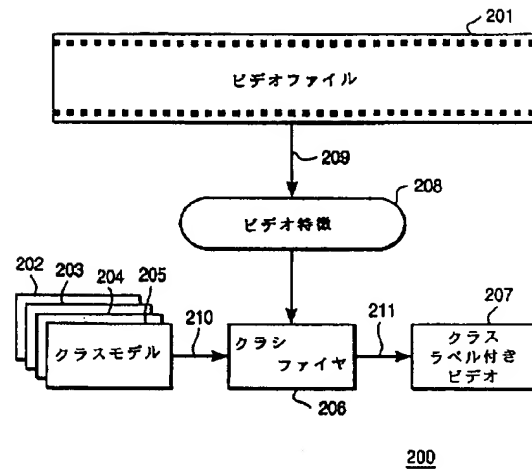
【図43】本発明によるスライド区間から抽出されたオーディオ区間からのソース特定話者モデルをトレーニングする方法を示すフローチャートである。

【図44】本発明による話者遷移モデルを用いたオーディオ・ビジュアル記録物をセグメント化する方法を示すフローチャートである。

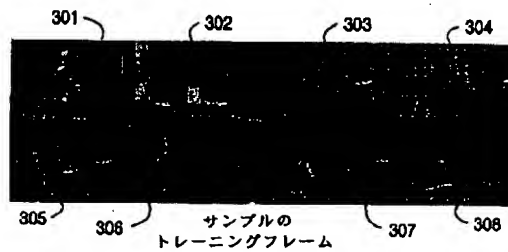
【図1】



【図2】



【図3】

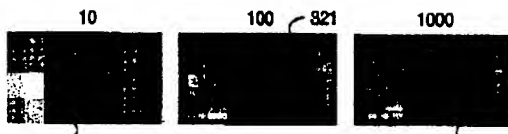


(A)



DCTガウスモデルの平均

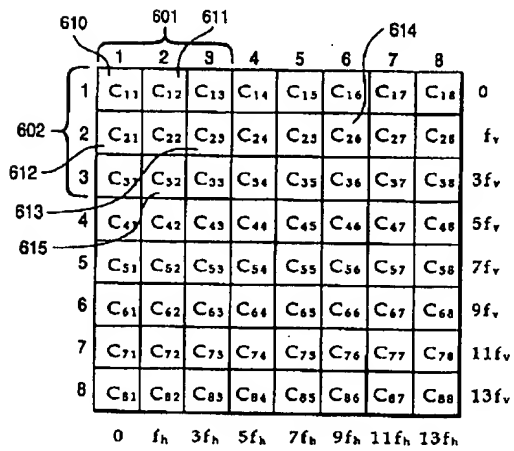
(B)



HTガウスモデルの平均

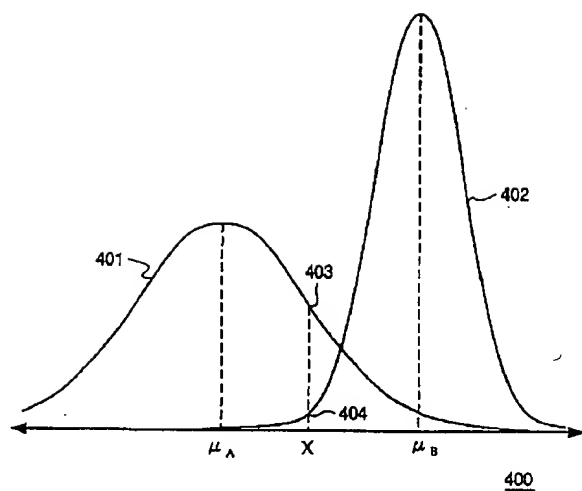
(C)

【図6】



600

【図4】



【図7】

	1	2	3	4	5	6	7	8
1	V <sub>11</sub>	V <sub>12</sub>	V <sub>13</sub>	V <sub>14</sub>	V <sub>15</sub>	V <sub>16</sub>	V <sub>17</sub>	V <sub>18</sub>
2	V <sub>21</sub>	V <sub>22</sub>	V <sub>23</sub>	V <sub>24</sub>	V <sub>25</sub>	V <sub>26</sub>	V <sub>27</sub>	V <sub>28</sub>
3	V <sub>31</sub>	V <sub>32</sub>	V <sub>33</sub>	V <sub>34</sub>	V <sub>35</sub>	V <sub>36</sub>	V <sub>37</sub>	V <sub>38</sub>
4	V <sub>41</sub>	V <sub>42</sub>	V <sub>43</sub>	V <sub>44</sub>	V <sub>45</sub>	V <sub>46</sub>	V <sub>47</sub>	V <sub>48</sub>
5	V <sub>51</sub>	V <sub>52</sub>	V <sub>53</sub>	V <sub>54</sub>	V <sub>55</sub>	V <sub>56</sub>	V <sub>57</sub>	V <sub>58</sub>
6	V <sub>61</sub>	V <sub>62</sub>	V <sub>63</sub>	V <sub>64</sub>	V <sub>65</sub>	V <sub>66</sub>	V <sub>67</sub>	V <sub>68</sub>
7	V <sub>71</sub>	V <sub>72</sub>	V <sub>73</sub>	V <sub>74</sub>	V <sub>75</sub>	V <sub>76</sub>	V <sub>77</sub>	V <sub>78</sub>
8	V <sub>81</sub>	V <sub>82</sub>	V <sub>83</sub>	V <sub>84</sub>	V <sub>85</sub>	V <sub>86</sub>	V <sub>87</sub>	V <sub>88</sub>

700

【図8】

1	C <sub>11</sub>	801
2	C <sub>12</sub>	802
3	C <sub>13</sub>	803
4	C <sub>21</sub>	804
5	C <sub>22</sub>	805
6	C <sub>23</sub>	806
7	C <sub>31</sub>	807
8	C <sub>32</sub>	808
9	C <sub>33</sub>	809

800

【図11】

【図9】

1	μ <sub>11</sub>
2	μ <sub>12</sub>
3	μ <sub>13</sub>
4	μ <sub>21</sub>
5	μ <sub>22</sub>
6	μ <sub>23</sub>
7	μ <sub>31</sub>
8	μ <sub>32</sub>
9	μ <sub>33</sub>

900

【図10】

	V <sub>11</sub>	0	0	0	0	0	0	0	0
0	0	V <sub>12</sub>	0	0	0	0	0	0	0
0	0	0	V <sub>13</sub>	0	0	0	0	0	0
0	0	0	0	V <sub>21</sub>	0	0	0	0	0
0	0	0	0	0	V <sub>22</sub>	0	0	0	0
0	0	0	0	0	0	V <sub>23</sub>	0	0	0
0	0	0	0	0	0	0	V <sub>31</sub>	0	0
0	0	0	0	0	0	0	0	V <sub>32</sub>	0
0	0	0	0	0	0	0	0	0	V <sub>33</sub>

1000

【図16】

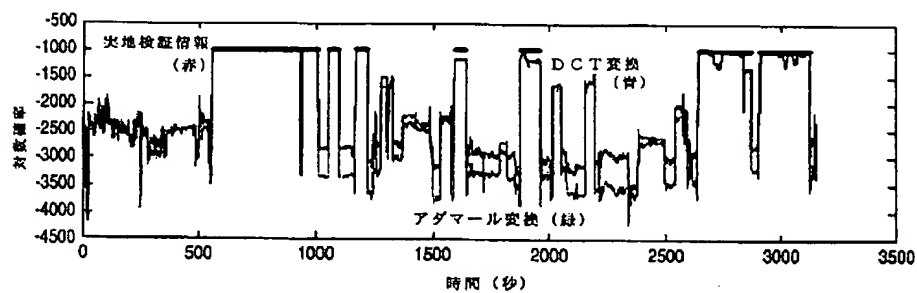
1	X <sub>11</sub>	1101
2	X <sub>12</sub>	1102
3	X <sub>13</sub>	1103
4	X <sub>21</sub>	1104
5	X <sub>22</sub>	1105
6	X <sub>23</sub>	1106
7	X <sub>31</sub>	1107
8	X <sub>32</sub>	1108
9	X <sub>33</sub>	1109

1100

1	X <sub>11</sub>	1601
2	X <sub>12</sub>	1602
3	X <sub>21</sub>	1603
4	X <sub>23</sub>	1604
5	X <sub>26</sub>	1605
6	X <sub>32</sub>	1606

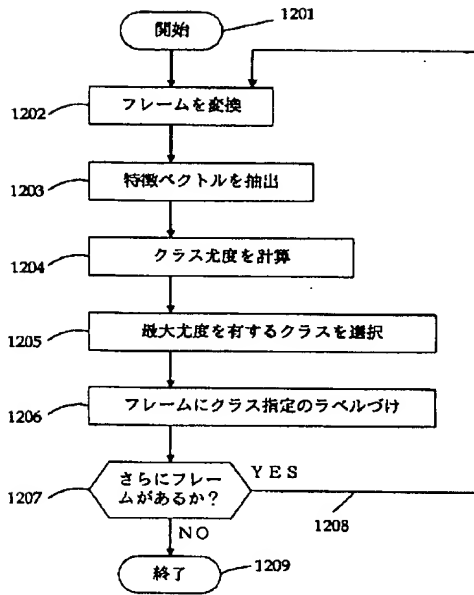
1600

【図19】



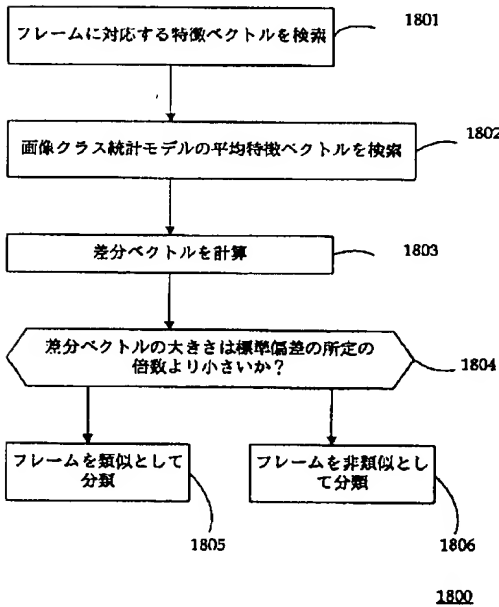
1900

【図12】



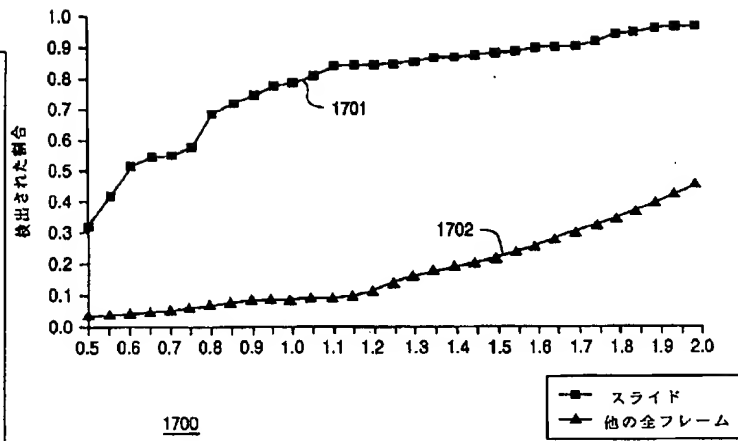
1200

【図18】

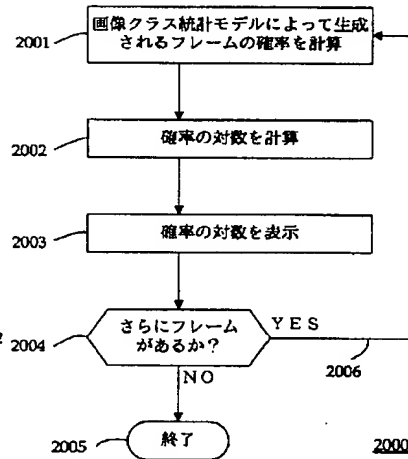


1800

【図17】

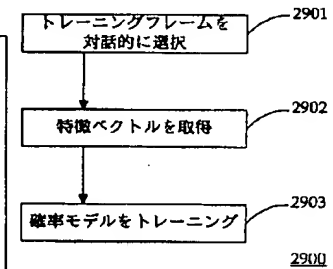


【図20】

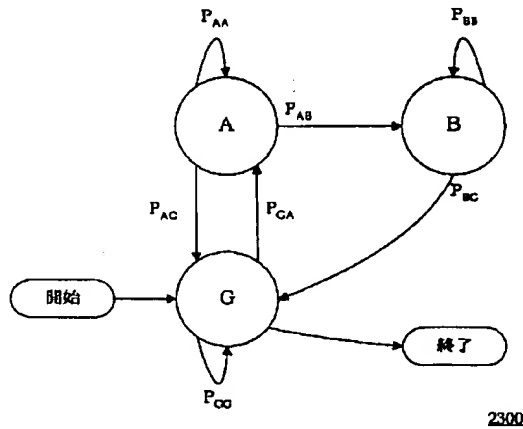


2000

【図29】

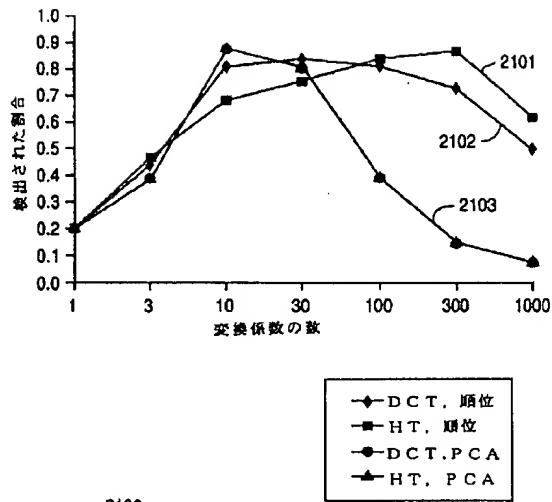


【図23】



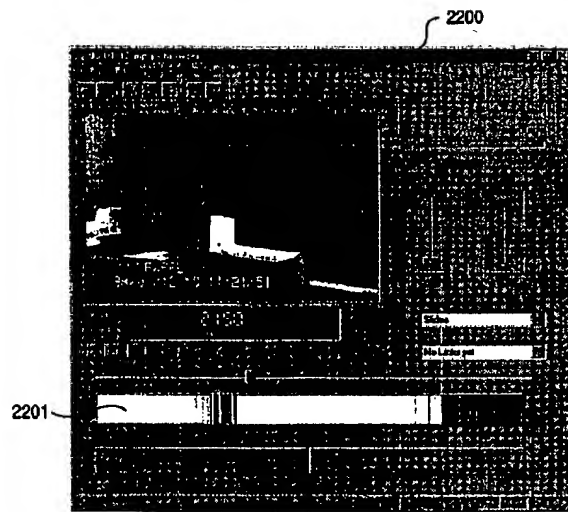
2300

【図21】

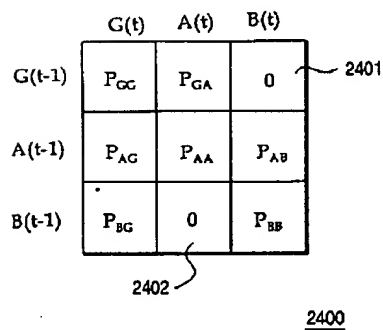


2100

【図22】

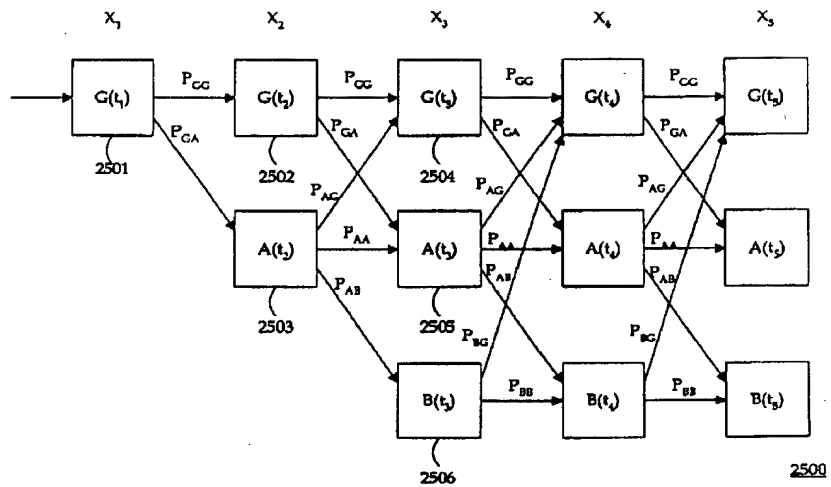


【図24】



2400

【図25】

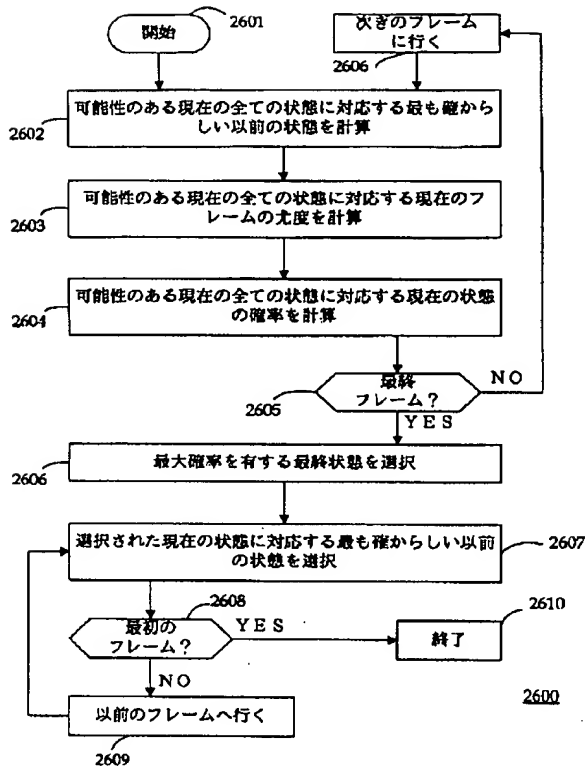


【図31】

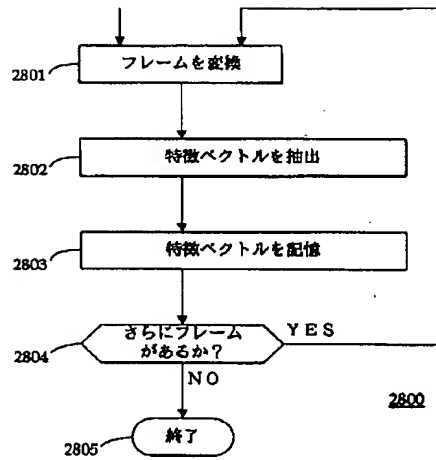


3100

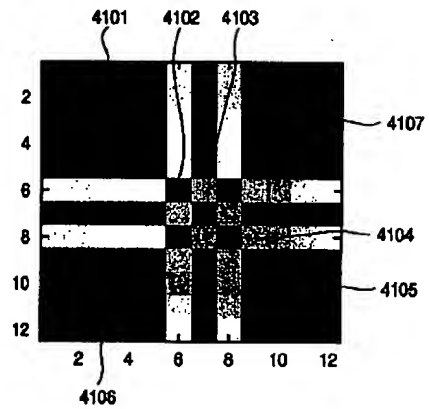
【図26】



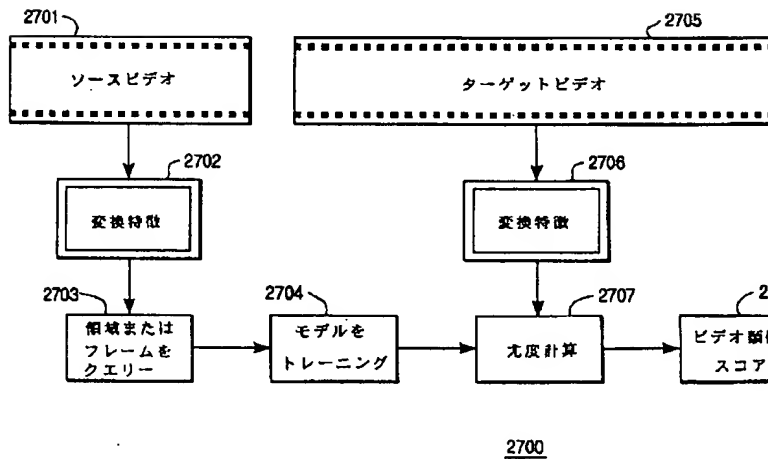
【図28】



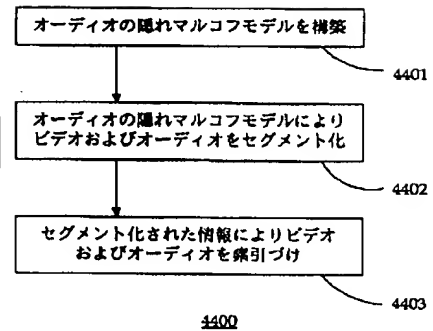
【図41】



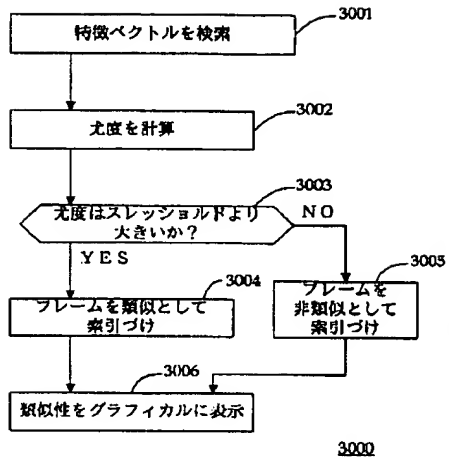
【図27】



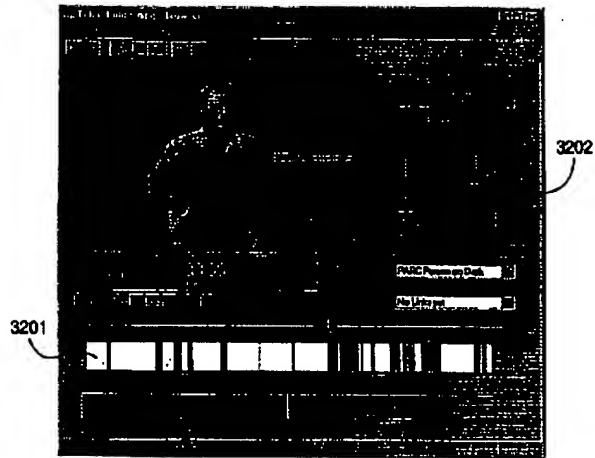
【図44】



【図30】

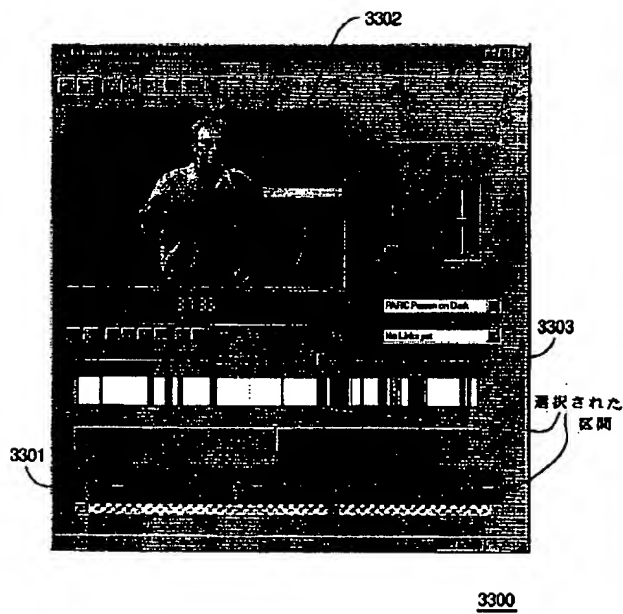


【図32】

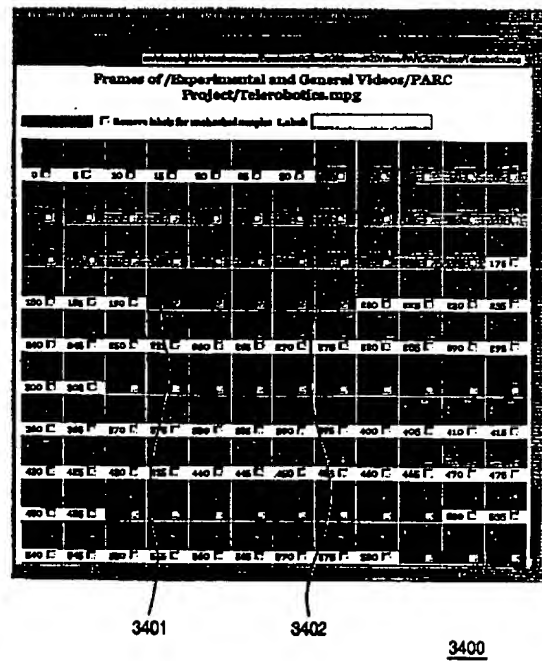


3200

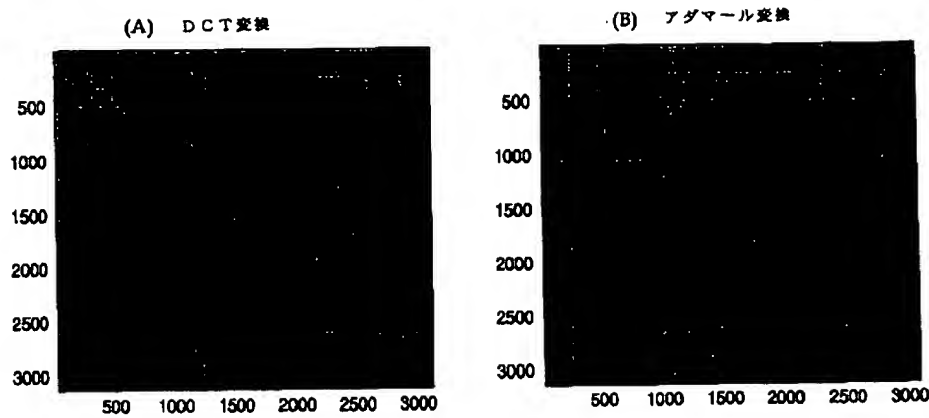
【図33】



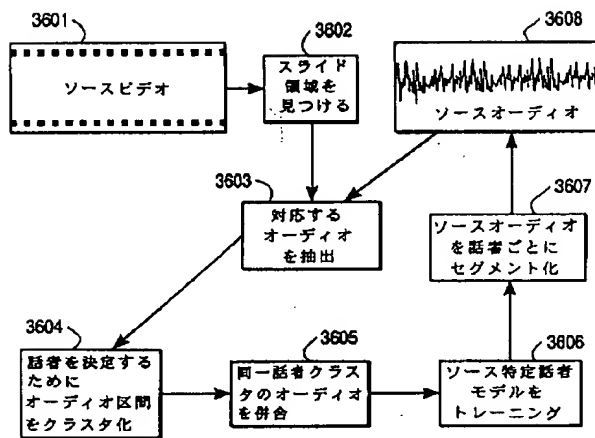
【図34】



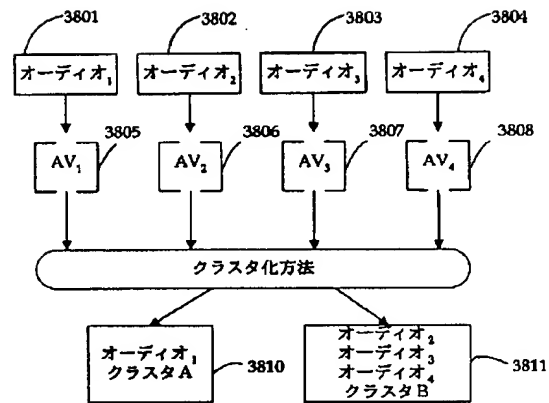
【図35】



【図36】

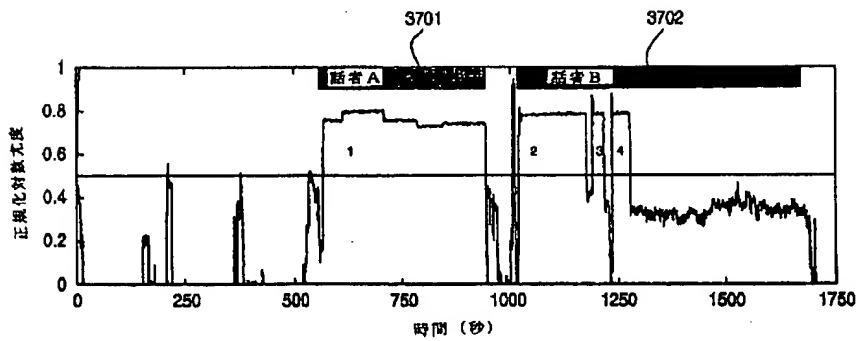


【図38】



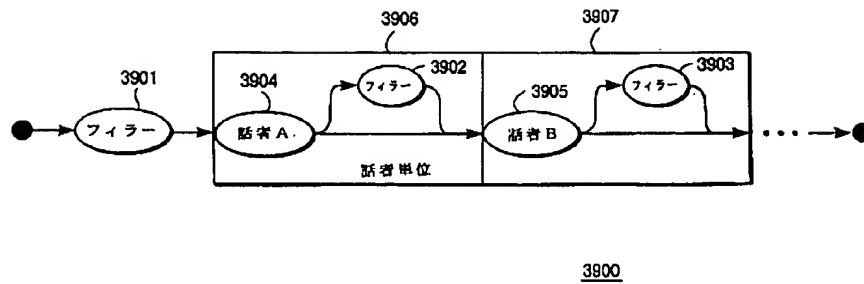
3800

【図37】

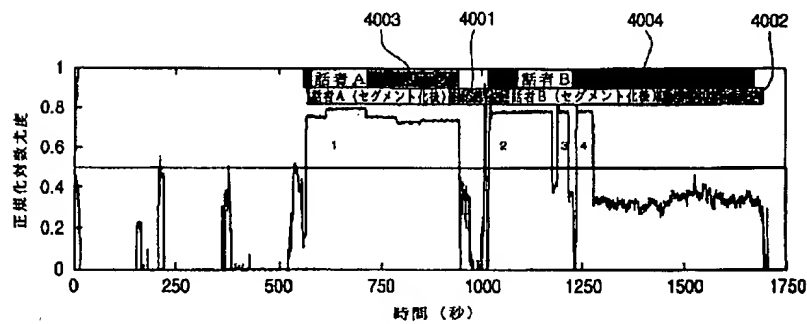


3700

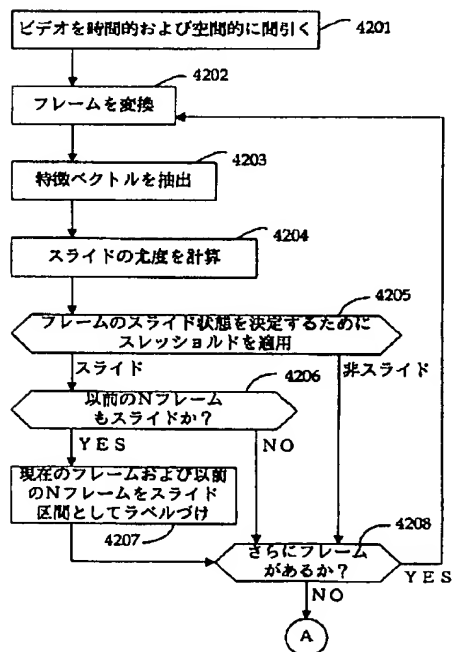
【図39】



【図40】



【図42】



【図43】

